Hybridizing Multiple Filters and GA Wrapper for Feature Selection of Microarray Cancer Datasets

Author

PAKIZAH

MS-17-CSE-00000205406

Supervisor

DR. USMAN QAMAR

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JANUARY, 2020

Hybridizing multiple filters and GA wrapper for feature selection of Microarray cancer datasets

Author

PAKIZAH

MS-17-CSE-00000205406

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Software Engineering

Thesis Supervisor:

DR. USMAN QAMAR

Thesis Supervisor's Signature: _____

DEPARTMENT OF COMPUTER AND SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

JANUARY, 2020

Declaration

I certify that this research work titled "*Hybridizing multiple filters and GA wrapper for feature selection of Microarray cancer datasets"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

PAKIZAH

MS-17-CSE-00000205406

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

PAKIZAH

MS-17-CSE-00000205406

Signature of Supervisor

Dr. Usman Qamar

Copyright Statement

Acknowledgements

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which Your setup in my mind to improve it. Indeed, I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but you.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life.

I would also like to express special thanks to my supervisor Dr. Usman Qamar for his help throughout my thesis and also for Data Mining and Web Engineering courses which he has taught me. I can safely say that I haven't learned any other engineering subject in such depth than the ones which he has taught.

I would also like to thank Dr. Wasi Haider Butt, and Dr. Saad Rehman for being on my thesis guidance and evaluation committee.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my exceptional parents and adored siblings whose tremendous support and cooperation led me to this wonderful accomplishment

Abstract

DNA Microarray technology is a valuable advancement in medical field but it gives birth to many challenges like curse of dimensionality, storage and computational requirements. Feature Selection is one way to handle these issues. To overcome the issues and challenges associated with microarray cancer dataset and not to compromise over relevancy, optimality and to improve the performance of metaheuristic Genetic Algorithm based wrappers, in this paper we have proposed, a multiple filters and GA wrapper based hybrid feature selection approach (MF-GARF) that incorporates Random forest as fitness evaluator of features. The proposed hybrid approach MF-GARF is comprised of three phases relevancy block; containing information theory based filters Information Gain, Gain Ratio and Gini Index, responsible for ensuring relevancy and removal of irrelevant and noisy features. Second phase is Redundancy block; incorporating Pearson Correlation statistics to remove redundancy among features, and then final phase Optimization Block; containing Genetic Algorithm wrapper with Random Forest as fitness evaluator, responsible for generating an optimal feature subset with high predictive power. Random Forest, kNN, Naïve Bayes and SVM within a 10-fold cross validation setup is used to calculate the classification accuracy of selected optimal feature subset. Experiments are carried out on 7 publically available benchmark binary and multiclass Microarray gene expression cancer datasets and the proposed algorithm has achieved good accuracy with minimal selected features for all datasets. The thorough comparison with other state of the art GA based and other metaheuristic hybrid techniques validates the effectiveness of our proposed approach in terms of features count and classification accuracy.

Key Words: *Genetic Algorithm, Microarray Gene Expression Datasets, Feature Selection, Information Gain, Gini Index, Gain Ratio, Correlation, Random Forest, Hybrid, Wrapper, Filter, Microarray Cancer Dataset, Gene Selection*

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

Data is too diverse. Diversity of data has made feature selection a fundamental step for many data mining tasks, especially for processing of high dimensional data like microarray datasets comprising of more than thousands of features and small set of samples. The rapid growth of data gives birth to many challenges like curse of dimensionality, storage and computational requirements. Gathering data is not a problem but obtaining meaningful information from raw data is critically important. The abundance of data demands optimal and efficient algorithms to process raw data to retrieve useful information [1] [2].

1.1 Microarray Technology

DNA Microarray technology [3][4] is a valuable advancement in medical field that facilitates medical specialists in monitoring and profiling gene expressions of an organisms. With the help of this technology, biologist can profile thousands of gene expressions in a single experiment. A microarray, also known as DNA chip, is basically a glass slide on to which DNA particles are fixed in a methodical way at specific areas called spots. A microarray may contain a large number of spots and each spot may contain million duplicates of indistinguishable DNA molecules that corresponds to a gene. Microarray Technology has several types e.g. Bacterial Artificial Chromosomes (BAC) microarrays, Oligonucleotide microarrays, Single Nucleotide Polymorphism (SNP) microarrays, and complementary DNA (cDNA) microarrays. Main application of microarray technology is profiling of cancer gene expressions. Firstly, RNA is extracted from the cell, then reverse transcription of extracted RNA molecules to cDNA is done with the help of the enzymes. Green, red, yellow and black fluorescent dyes are used to label cell samples satisfying particular conditions (for say, red for tumor and green for normal) and then hybridized on a single glass. After this, hybridized microarrays are scanned at suitable wavelengths, and finally an image is generated which is processed to obtain gene expression data.

DNA chip can profile thousands of genes, but not all gene expressions contribute to the diagnosis of a disease. Microarray gene expression dataset contains plenty of irrelevant and redundant genes that may halt the process of correct diagnosis of a disease [4].

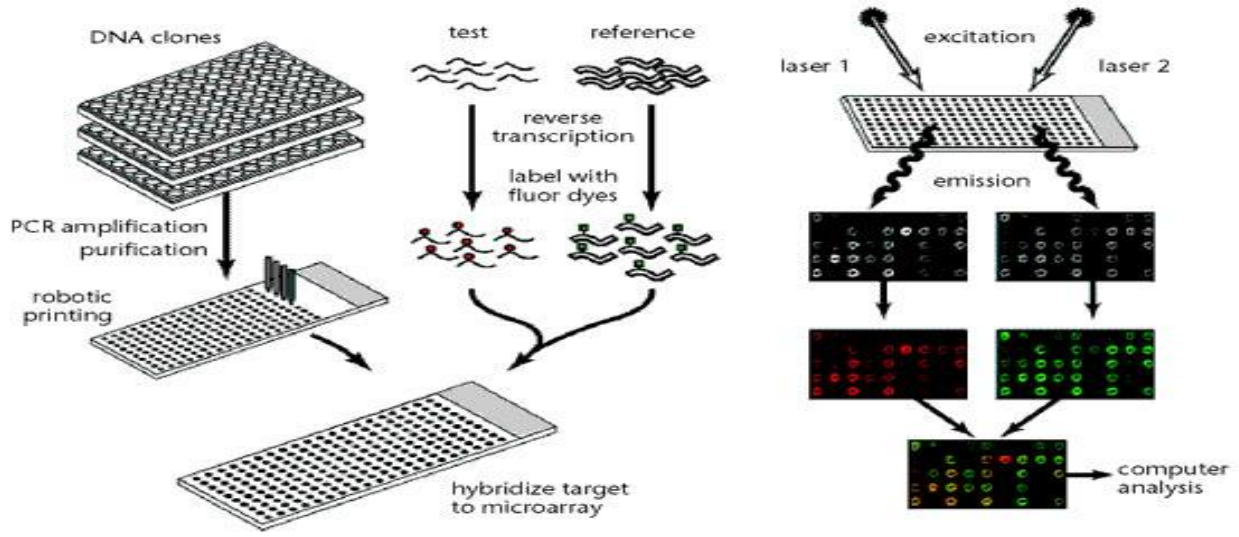

Figure 1.1-1: Process of Microarray Genes Expression Profiling

1.2 Microarray Gene Expression Analysis

The process of analyzing the gene expression dataset to find out the most informative genes from the pool of noisy, redundant and irrelevant is known as Microarray Gene expression analysis [4]. Analysis of microarray gene expression datasets is a crucial task to resolve the issues and challenges associated with them. Moreover, it helps the biologist to tackle things and interpret microarray datasets at molecular level. Classification of microarray gene expressions is a NP hard problem [3], it contains thousands of genes and small sample size, that gives birth to the issue of curse of dimensionality. To deal with the issue of high dimensionality and low sparsity, dimensionality reduction is one solution [5]. Dimensionality reduction [6] in terms of feature (gene) selection is of great interest as large number of features (genes) and small sample size leads to overfitting of model, poor model learning, erroneous predictions. Moreover, model construction and learning over such dataset are computationally expensive and inefficient.

In data sciences, feature selection [7][8] is one of the most important concept. Feature selection aims to select the feature subset from the original feature set based on feature relevancy and redundancy. The selection of good and most relevant features directly influences the performance of training model. The training of model on relevant features positively impacts the accuracy of model while it impacts negatively if learning of model is carried out on irrelevant or partially relevant features.

The process of selection of features that contribute more to the accurate prediction can be performed manually or automatically using statistical and model based learning approaches. Generally, Feature selection strategies are classified into four categories, filters, wrapper, embedded and hybrid. Filters select features by evaluating each feature individually and scoring statistically without using the heuristics of any classifier. Wrapper evaluate features using performance accuracy of the classifiers and are more efficient in terms of performance than filter but are computationally expensive. Embedded approach embeds feature selection in learning algorithm. Hybrid is a combination of any of the two or more feature selection approaches to overcome the issues associated with the individual approach and merge the goodness of the combined approaches.

1.3 Problem Definition and Challenges

Feature selection of microarray gene expression dataset also known as Gene selection as a research domain, poses many serious challenges.

1. The challenges originate from the exclusive natural environment of the prevailing gene expression dataset; where almost all of these datasets have instances below 200 against hundreds and thousands of features.
2. There are only a few features (genes) among these thousands of genes that are relevant to predictive class.
3. The noise is inherent in these gene expression datasets due to natural or technical reasons.
4. Another challenge is associated with the application area, for say, the accuracy is an important criterion in cancer classification process. But, in cancer domain we don't have to just achieve accuracy, but biological relevancy and reliability too.
5. Traditional classifier lacks the capability to classify high dimensional datasets.

1.4 Motivation

Microarray gene expression dataset contains thousands of genes and few instances. The high dimensionality and low sparsity of Microarray datasets pose serious challenges for data mining. Not all features (genes) contribute to the prediction of a disease. Such dataset contains

noisy and irrelevant features that negatively impact the model training and ultimately the prediction accuracy of the standard classifiers. The processing of such kind of data and extracting useful information from it, is a tedious task. Challenges associated with microarray cancer datasets [5] are curse of dimensionality, low sparsity, noise, redundancy and irrelevancy. Standard classifiers lack capabilities to deal with these challenges all in all and fail to classify the correct classes. Feature selection is an effective way to solve these issues of microarray datasets. It is a common preprocessing techniques in data mining tasks to enhance the efficiency of classifiers. Major concern in feature selection is to select a subset of features incorporating maximum relevancy and minimum redundancy. The analysis and classification of microarray dataset are an active research area in biomedical field, specially there is great demand for novel and reliable approaches of feature (gene) selection for microarray cancer dataset to classify the deadly diseases.

This thesis presents a novel hybrid approach based on multiple tree based filters and GA wrapper with an aim to create a small informative and optimal feature pool for classification of microarray cancer datasets free from noisy, irrelevant and redundant features. At stage one, for relevancy analysis of microarray cancer gene expression datasets tree based filters i.e. Information gain [9], Gain ratio [10], and Gini index [11] are used. In second stage, the Correlation filter [12] is used to remove the redundancy from set of informative genes. And then in final stage, further refinement of feature pool into an optimal subset is carried out using a promising bio-inspired evolutionary Genetic Algorithm [13] wrapper which achieves best prediction accuracy. Each Filter technique has its own plus points and weaknesses. But relying on just one filter technique results in biasness. So in this hybrid approach we are incorporating features ranked by each filter into a unified feature set meeting a specific threshold criterion and then pass it to the starting space of GA wrapper for refinement and optimization.

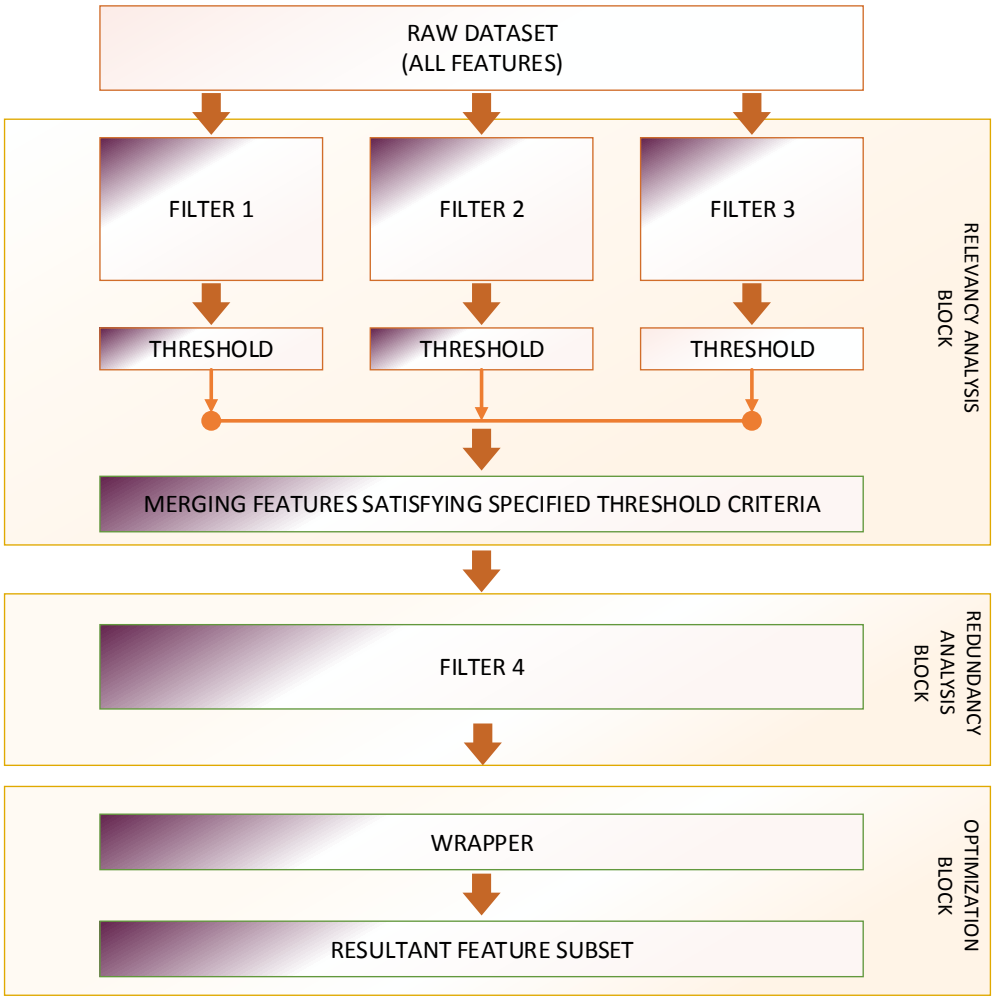

Figure 1.4-1: Block Diagram of Proposed Hybrid Approach

1.5 Thesis Objective and Contribution

The main objective of this study is to achieve an optimal subset for microarray cancer dataset with high predictive power, so they can identify the cancer type. Secondly improve the performance of Genetic Algorithm based wrapper in terms of classification accuracy.

This thesis presents a novel hybrid approach based on multiple information theory based filters and Genetic Algorithm based wrapper with an aim to create a small informative and optimal feature pool for classification of microarray cancer datasets free from noisy, irrelevant and redundant features.

The introduction of multiple filters as a preprocessor, improves the performance of Genetic Algorithm based wrapper for high dimensional Microarray Gene Expression Dataset. And finally the selected features (genes) boost the performance of traditional classifiers too.

1.6 Thesis Structure

The structure of thesis is as follow. The chapter 2 discusses about the feature selection approaches. Chapter 3 provides a detailed literature review of existing feature selection approaches and analyses the gaps in literature. Chapter 4 gives the overview of proposed methodology, the techniques and classifier it has employed, implementation details and the flow chart to depict the flow of proposed methodology. Chapter 5 covers the experimental setup, datasets and results. Chapter 6 presents the conclusion and suggest future direction.

CHAPTER 2: DIMENSIOALITY REDUCTION

There are two main approaches for dimensionality reduction, feature extraction and feature selection. Feature Extraction [5] is also known as transformation; it transforms the original feature set into a new reduced feature set comprising of meaningful information based on combination of original features. It's an effective approach for dimensionality reduction but lack interpretability. Principal Component Analysis [14], Latent semantic analysis [15], Linear Discriminant Analysis [15], Independent component analysis [17] etc. are some examples of feature extraction techniques.

Feature Selection [5], on the other hand, is a process of selecting a feature subset from original feature set comprising of most relevant information with respect to the target class. It does not involve transformation thus retain the interpretability and originality of each selected feature, unlike feature extraction. Information Gain[9], Gini Index[10], Chi Squares[11], Correlation[12], Genetic Algorithm[9][13], Lasso[18] etc. are some example of feature selection techniques.

In this section we cover the overview of feature selection, stages of feature selection and feature selection models.

2.1 Overview of Feature Selection

In data sciences, feature selection is one of the most important concept. Feature selection aims to select the feature subset from the original feature set based on feature relevancy and redundancy. The selection of good and most relevant features directly influences the performance of training model. The training of model on relevant features positively impacts the accuracy of model while it impacts negatively if learning of model is carried out on irrelevant or partially relevant features.

Traditionally, mostly feature selection techniques work to look for relevant features leaving redundant features unattended thus negatively impacting model leaning process. Yu and Liu [19] has redefined traditional framework for feature selection and improvised three disjoint categorizes of features into four categories. They have involved redundancy as a considerable trait of feature to ponder over. The categories are: (1) irrelevant features, (2) weakly relevant and

redundant features, (3) weakly relevant but non-redundant features, and (4) strongly relevant features.

Irrelevant and noisy features are the least important and worthless features containing no useful information, thus do not contribute to the predicting power of model. Weakly relevant feature can play an effective role but only if they are non-redundant, because redundant features being highly correlated to each other and having similar ranking disturb the target distribution. So it's better to remove irrelevant and redundant feature. Strongly relevant features play the most influential role in classification. They positively affect the discriminative power and prediction accuracy of model. Optimal feature subset is usually composed of features falling in last two categories. Therefore, in order to build a good predicting model, target should be to come up with a feature subset containing strongly relevant feature and non-redundant-weakly relevant features, eliminating noisy and irrelevant features. Fig. 3 shows feature categorization based on relevancy and redundancy.

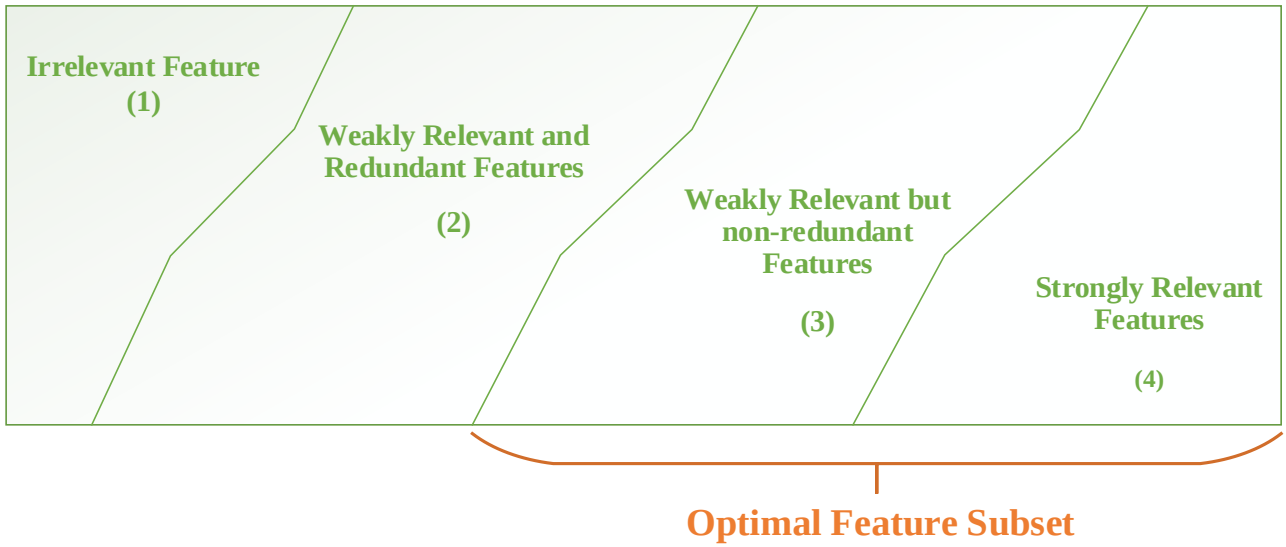

Figure 2.1-1: Categories of Features

.

Yu and Liu [19] have defined feature redundancy formally, and introduced redundancy analysis in feature selection traditional framework analyzing the relationship of feature relevancy and redundancy.

Thus, Feature selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. Motivation of this study is to create a small

feature pool using fast filters considering both redundancy and relevancy analysis, further refine the feature pool into an optimal subset using promising wrapper to achieve better prediction accuracy.

2.2 Stages of Feature Selection

Feature selection [5] is a process of searching a subset from original set of features incorporating highly informative features with maximum relevancy to the target class and minimum redundancy. And this process of selection is comprised of four main stages as shown in the figure 2.2-1. And every stage carries an influence on overall performance of feature selection process.

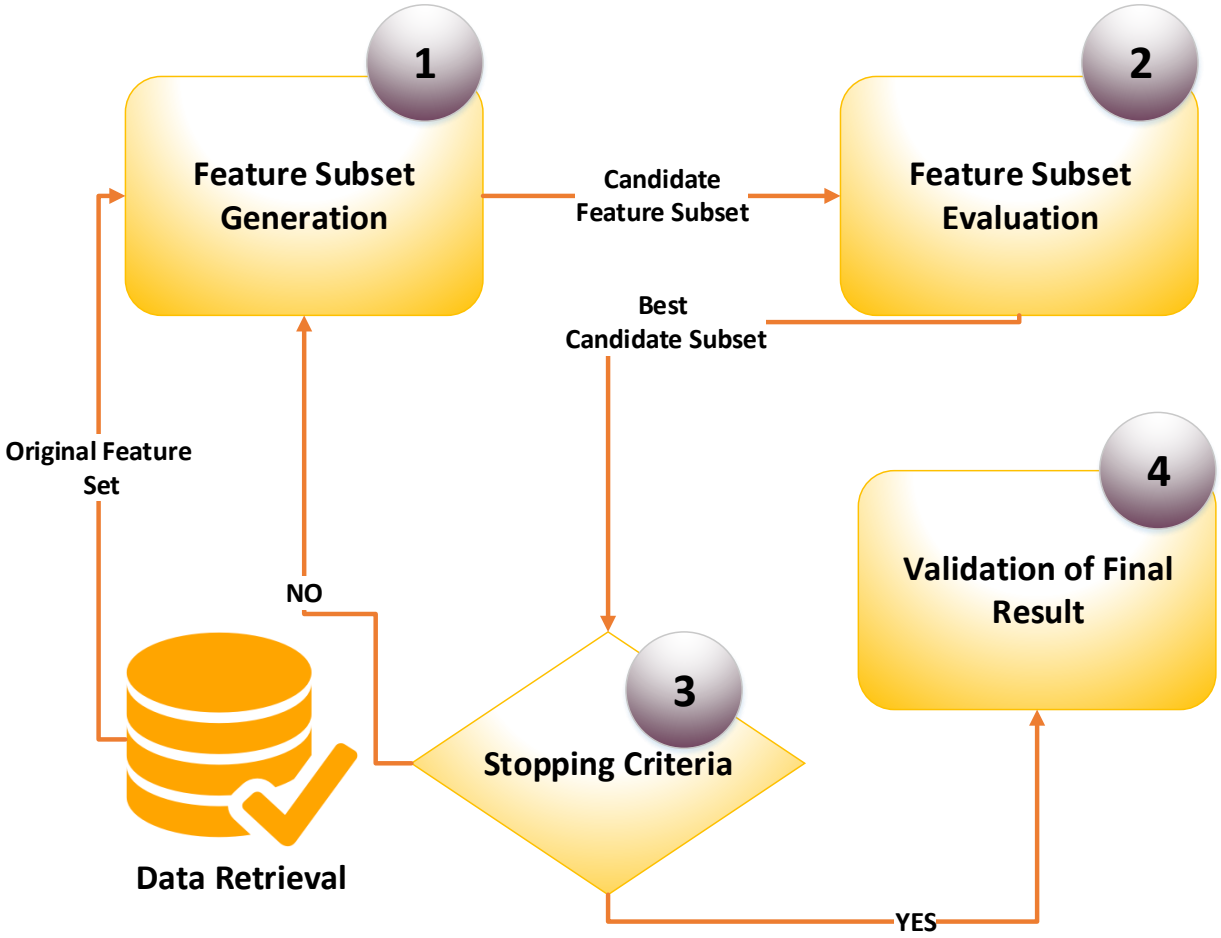

Figure 2.2-1: Stages of Feature Selection Process

2.3 Feature Selection Strategies

Generally, Feature evaluation strategies are classified into four categories, filters, wrapper, embedded and hybrid. But recently, a new feature selection strategy is developed, called ensemble feature selection.

2.3.1 Filters

Filters [20] are the most commonly used features selection strategy that selects feature based on the intrinsic properties of data without any direct involvement of learning algorithms or classifiers. Filter are also known as open loop methods. Filter algorithms measure the feature traits using typically four types of evaluation functions based on information, distance, dependence, and consistency and rank the features accordingly. Filter score features statistically independent of any classifier that's why they are computationally inexpensive and more efficient. And easy to scale up for high dimensional datasets. Being independent of any learning algorithm, filters methods are unbiased and provide a generalized resultant output for all kind of classifiers.

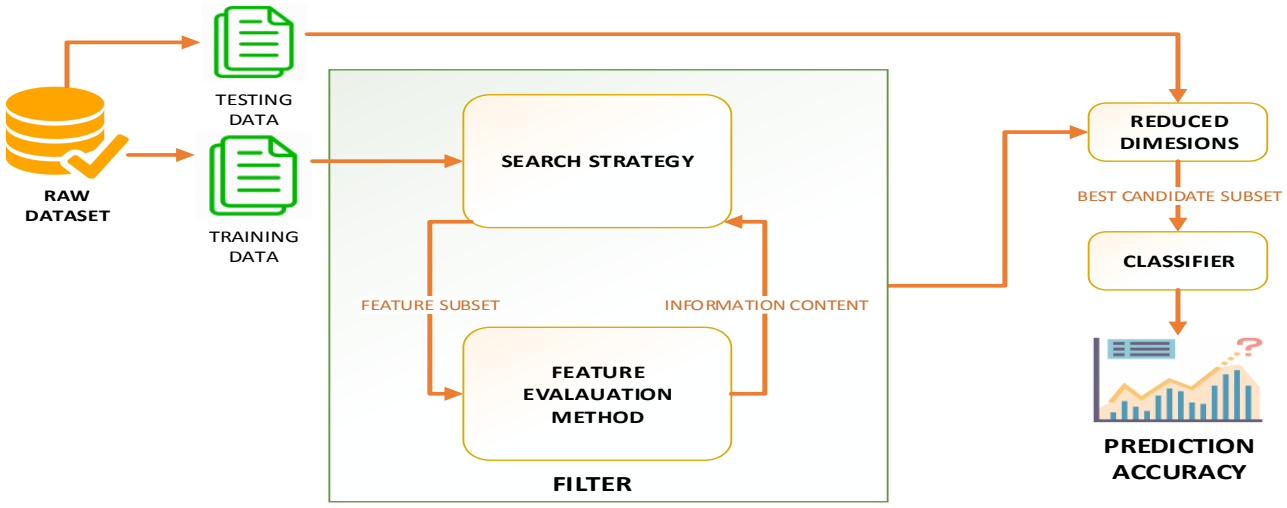

Figure 2.3-1: Framework of Filter Approach

There exist so many feature ranking and selection methods. There are two broad categories of filter methods.

- Univariate Filters

- Multivariate Filters

The univariate filter uses feature evaluation function (FEF) for scoring each feature considering ability of features to differentiate different classes from each other. Few examples of feature evaluation functions are Information gain, Mutual Information [25], Chi Square [26] and Gini Index.

Major issue with FEF's resultant feature subset is that features despite of being highly scored are redundant with respect to each other. Sometime it happens that resultant feature subset, though have highly scored features, fails to separate the overlapped classes while low scored feature despite of having capability to separate overlapped classes fail to make their mark in resultant feature subset because of their low scores.

Univariate filters evaluate the relevance of features independently while Multivariate filter, on the other hand, perform better by handling not only relevancy but redundancy by considering the correlation among the features. But, multivariate filters are computationally expensive than univariate filters. Examples of multivariate filters are correlation based feature selection, and fast correlation based feature selection.

2.3.2 Wrapper

Wrapper approach [21], also known as a closed loop method, is an efficient feature selection strategy. It performs feature selection considering the performance evaluation by learning algorithm or classifier. Wrappers are generally categorized as deterministic and randomized wrappers. Randomized wrapper approaches generally generate better results but are computationally expensive than deterministic wrapper approaches. Few examples of wrapper approaches are Forward elimination, Backward elimination, Genetic Algorithm and Beam Search Algorithm.

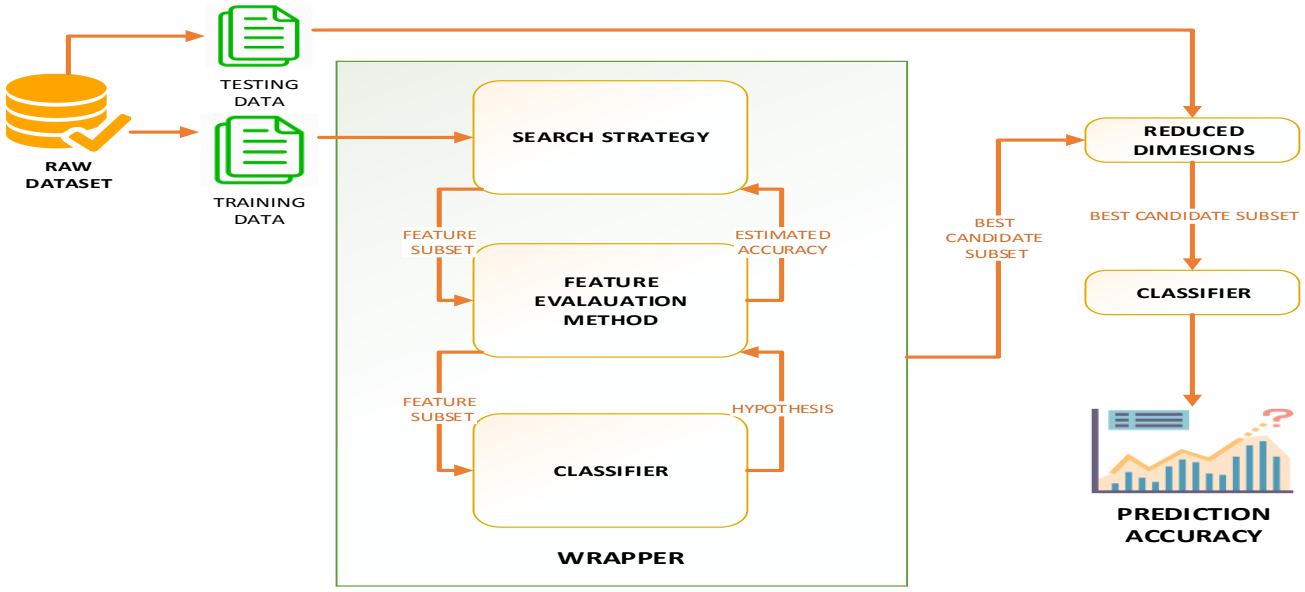

Figure 2.3-2: Framework of Wrapper Approach

2.3.3 Embedded

Embedded [22] approach is quite similar to the wrapper approach in performance but comparatively more computationally tractable, it embeds feature selection within induction algorithm and uses its properties as an evaluation function. It is less prone to overfitting as compared to wrapper but computationally expensive for high dimensional datasets. It also considers the dependencies among the features but specific to learning to algorithm.

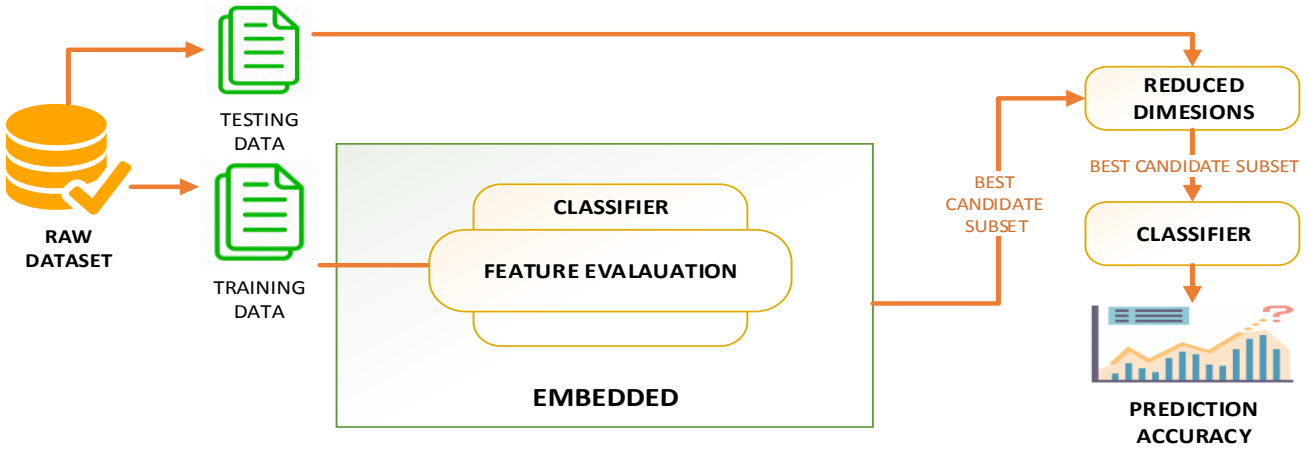

Figure 2.3-3: Framework of Embedded Approach

2.3.4 Hybrid

Hybrid is generally a combination of two or more feature selection strategies. It can either be combination of more than one filters or a filter-wrapper, filter-embedded or wrapper-embedded approaches. The most common hybrid method in literature is of filter- wrapper combination [23]. The advantage of hybrid over other techniques is that it merges the advantages or positive aspects of two or more different techniques thus overshadowing the negative aspects of these incorporated strategies. For say, in a filter-wrapper combination, using filters as a preprocessor for removal of noisy and irrelevant features, minimize the initial search space for wrapper and improves performance and overall computational cost.

2.3.5 Ensemble

Ensemble approach [24] is new to the court of feature selection strategies. It does not depend on the feature ranking done by only one feature evaluator instead ranking of features is finalized by aggregating the ranking given by multiple feature selectors. It is developed to overcome the instability issues of different feature selection algorithm. It's a good option for high dimensional dataset as it generates a stable feature subset, but again as this technique is dependent on multiple feature selectors so it would be quite computationally expensive.

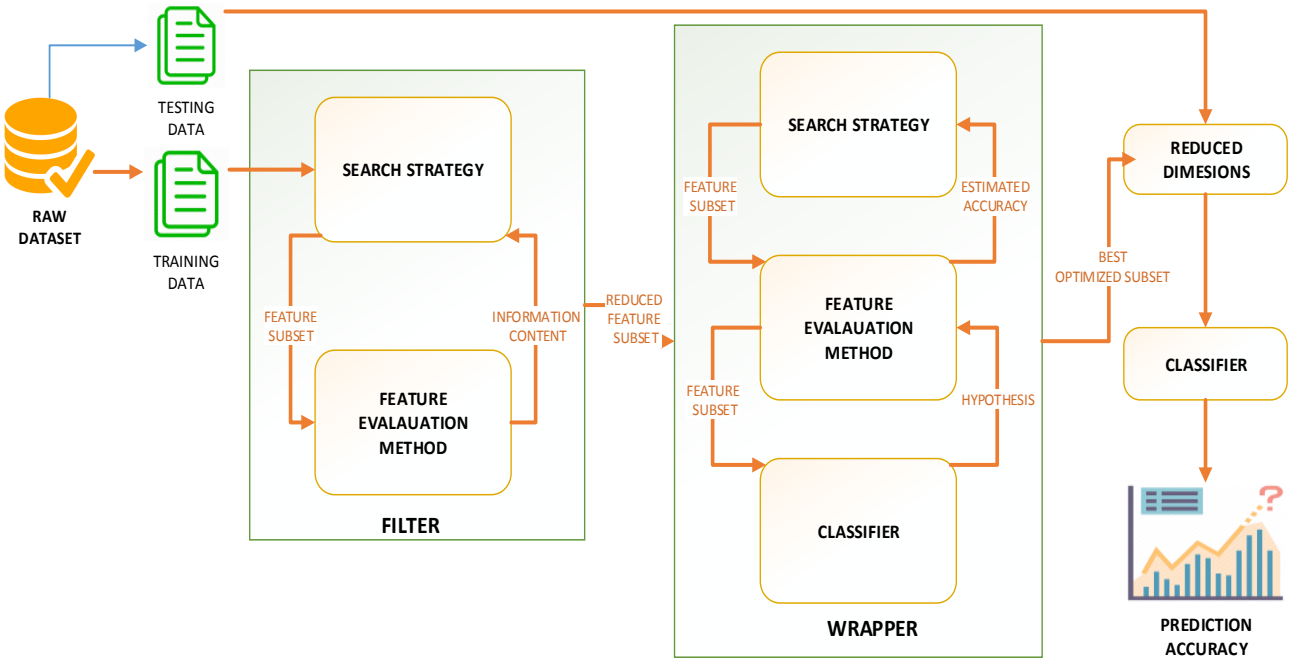

Figure 2.3-4: Framework of Hybrid Approach

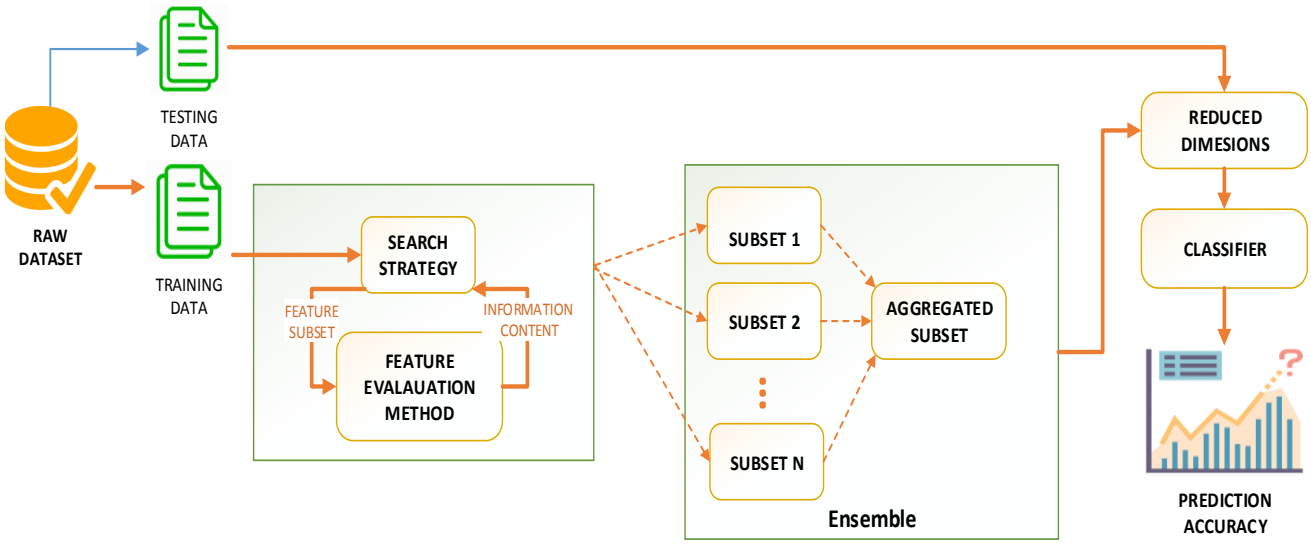

Figure 2.3-5: Framework of Ensemble Approach

Each feature selection strategy has its own pros and cons, but the main aim of feature selection is to select a subset of features incorporating maximum relevancy and minimum redundancy.

This is because[19]:

a) irrelevant features do not have discriminative power, play no role in model learning performance, thus removing irrelevant feature ensures better model performance.

b) redundant features that might be relevant but repetitive, do not contribute any additional information to model thus removing one of them do not affect the model learning process.

The table 2.3-1 presents a summary of characteristics of all feature selection methods. All feature selection methods are classifier dependent except filter methods, that's why filter methods are fast and computationally inexpensive in comparison to other techniques. Wrapper method are highly computational expensive than filters, that's the reason hybrid approaches are preferred over wrappers to overcome the risk of overfitting and computational cost. All feature selection methods except filter consider feature interaction that why their resultant feature are more stable.

Table 2.3-1 : Characteristics of Feature Selection Strategies

		Characteristics					
		Classifier Dependence (Yes/No)	Account Feature Dependence (Yes/No)	Computationally Intensive (Less/ Highly/ Moderately)	Risk of Overfitting (Low/Medium/ High)	Scalable (Yes/No)	Coverage
Feature Selection Strategies	Filter	No	No	Less	Low	Yes	Relevancy
	Wrapper	Yes	Yes	Highly	High	No	Relevancy and Redundancy
	Embedded	Yes (Highly Dependent)	Yes	Highly (Better than wrappers)	Medium	No	Relevancy and Redundancy
	Hybrid	Yes	Yes	Moderately	Low	No	Features Optimality
	Ensemble	Yes	Yes	Highly	Low	No	Reliability and Stability

CHAPTER 3: LITERATURE REVIEW

DNA Microarray technology (DNA-chip) has made collection of gene expression data much easier. It can profile thousands of gene expressions in single experiment. Nowadays, the DNA-Chip, an advanced technology is widely used in biomedical research area for profiling cancer gene expression data [3]. But, advancement in technology such as next generation sequencing, mass spectrometry and microarrays, gives birth to plenty of serious challenges. Many studies have analyzed these gene expression datasets for classification of diseases and to identify its subtypes using various data mining techniques. Microarray datasets have large dimensions and low sparsity issues due to which basic data mining algorithms do not perform well and fail to predict disease accurately.

Problem Associated with Microarray Datasets are as follow: -

- Microarray Gene Expression Dataset is a high dimensional data, it contains thousands of feature and few instances. That results in curse of dimensionality issue. That can lead to overfitting of machine learning algorithms.
- The biggest issue associated with microarray dataset set is irrelevant and redundant features that effect the discriminative powers of features, impacts model training and ultimately results in bad classification accuracies.
- Limited number of instances in microarray dataset are in un-balanced state i.e. instances are not equally divided among each class, effects training and testing data splits.

When we have high dimensional datasets the main challenge is to select a feature subset that is reliable and has the discriminative powers to correctly predict the target class. In an attempt to overcome these challenges and issues, dimensionality reduction and feature (gene) selection algorithm have been applied. In this section we have discussed the existing state of the art filter, wrapper and hybrid approaches for supervised feature selection of microarray caner datasets. For that purpose, we have focused on the literature of recent 10 years spanning from 2010 to 2019. And the databases we have searched for literature include IEEE, Springer, Science Direct, Hindawi, and ACM.

3.1 Filters approaches for feature selection

Filter techniques because of it generalization properties have been used by many researchers for feature selection of microarray cancer datasets. The recent researches has presented many novel filter approaches to rank features like Hidden Markov's Model (HMM) [27], X- variance [28], Mutual congestion [28], Qualitative Mutual Information [29] used Mutual information (MI) with Random forest feature importance, Pareto based feature Ranking technique [30] for multi-objective optimization, Partial Maximum Correlation Information (PMCI) [31], a multiple synergy filter based feature selection approach that assesses feature importance by extracting orthogonal components from feature space. Correlation based feature selection [32], mRMR [33] approach that covers both the relevancy and redundancy in parallel manner.

Table 3.1-1: Description of Reviewed Filter Approaches for Feature (Genes) Selection

Literature	Proposed Filter Gene selection approach	Classifier	Microarray Cancer Datasets	Performance Validation Criteria
[27]	Hidden Markov's Model (HMM)	▪ SVM	▪ Prostate ▪ DLBCL ▪ Leukemia	LOOCV
[28]	X-Variance / Mutual Congestion	▪ NB ▪ DT ▪ SVM	▪ Colon ▪ CNS ▪ Leukemia1 ▪ Leukemia2 ▪ DLBCL ▪ Prostate	Accuracy, Specificity, Sensitivity
[29]	Qualitative Mutual Information (QMI)	▪ NB ▪ C4.5 ▪ IB1	▪ Breast ▪ Colon ▪ Leukemia	Classification Accuracies
[30]	Pareto Based Feature Ranking	▪ NB ▪ kNN ▪ ANN	▪ Leukemia ▪ CNS ▪ Prostate ▪ Lung ▪ Colon ▪ DLBCL	Accuracy, Specificity, Sensitivity

[31]	Partial Maximum Correlation Information (PMCI)	▪ SVM ▪ RF	▪ GLI ▪ Prostate(GE) ▪ Leukemia	Kappa, Jackknife CV
[32]	Correlation Based Feature Selection	▪ C4.5 ▪ CART	▪ Leukemia ▪ High Grade Glioma Dataset	Classification Accuracies
[33]	Improved Maximum Relevance Minimum Redundancy (mRMR)	—	▪ Prostate ▪ Childhood Leukemia ▪ Ovarian Cancer ▪ Leukemia	10 fold cross validation, AUC, F-Score, Average Correlation, Accuracy, Specificity, and Sensitivity
[34]	IG-TR CCA-TR	▪ Resilient Propagation ▪ Back Propagation ▪ Manhattan Propagation ▪ SVM	▪ Leukemia ▪ Medulloblastoma ▪ Lymphoma ▪ Prostate	KSI, BCR, BER

In this table 3.1-1 we have covered literature review of recent filter feature selection methods for microarray gene expression datasets, these recent techniques are improved version of existing feature ranking methods or few are combinations of multiple feature ranking methods [27] [30] [34] are better than traditional filter approaches. Momenzadeh et al [27] topology of HMM combined five feature ranking methods ROC, t-Test, Bhattacharya distance, Wilcoxon Test and entropy. The proposed model performed better than the individual feature feature ranking techniques mentioned earlier. Another paper [30] proposed a Pareto based feature Ranking technique for multi-objective optimization, it involved 7 ranking techniques Signal to Noise Ratio (SNR), Information Gain (IG), Pearson Correlation Coefficient (PCC), Kruskal-Wallis test, Fisher- Score, t-Test and Relief-F to create Pareto model to rank features and perform gene selection. This method has not relied on just one feature ranking method instead considered 7 ranking methods. At once the proposed filter technique use combination of only 2 out of 7 ranking criteria to create 21 Pareto Ranking Model and evaluated these model using 3 standard classifiers k- Nearest Neighbors (kNN), NB and Artificial Neural Network (ANN).

Alirezanejad et al paper [28] present two filter approaches X-Variance and Mutual Congestion to improve the performance of 3 standard classifiers Naïve Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM) for classification of biomedical datasets. X-variance rank features by considering the internal attributes of features while Mutual congestion rank features on the basis of frequencies of instances in intersection region of features. The experiments were carried out on 8 biomedical datasets including 6 high dimensional microarray datasets. Experiments conclude that Mutual Congestion performs better for high dimensional datasets while for low dimension, X variance is a better option. Accuracies achieved for microarray datasets were within a range of (70%-90%) which are definitely not exception for the bioinformatics datasets. Moreover, in this paper the researcher has also tried both filters in combination as XV- MC and MC-XV but results were not satisfactory.

[29] proposed a filter approach called Qualitative Mutual Information for Gene selection. The proposed filter approach combined information-based filter approach mutual information (MI) with Random forest feature importance to rank each feature and then assessed the quality of selected features using three standard classifiers Naïve Bayes (NB), C4.5, Instance Based Classifier (IB1). Accuracies achieved by proposed filter were good but no. of selected feature were too large.

All these filters are most recent approach and much better than traditional feature ranking approaches but still they lack the capabilities to produces lesser number of features when it comes to microarray dataset, and for most of these techniques classification accuracies range between 70%-90% and features count was around 100.

Filter based Feature selection has more generalization properties as compared to other approaches but they lack the capabilities to reduce the dimensions in case of high dimensional datasets [28] and thus do not generate the good prediction accuracies [29]. To overcome the drawbacks associated with filter approaches, wrapper and hybrid approaches are proposed that involve the heuristics of classifiers to evaluate the performance of selected features.

3.2 Wrapper Approaches for Feature Selection

In wrapper approach, search process is wrapped around an induction algorithm usually a classifier. And, the performance accuracy or classification error rate of the classifier is used to

evaluate the selection of best feature subset. Wrapper are more efficient than filters in terms of performance as it considers the correlation among features and directly incorporates the biases of induction algorithm. Many studies has proposed wrapper based approaches for feature selection of microarray cancer datasets like PSO-SVM [35], GA- SVM [35], ABC-SVM [35], FF-SVM [36], HS-GA [37], ACO-SVM [38], BPSO-CGA [39], and HPSO-LS [40]. Wrapper methods are better alternative to filters for supervised learning problems being efficient in performance but are computationally expensive, hence require plenty of computational resources for high dimensional datasets. Moreover, wrapper models are prone to overfitting, calling classifier again and again for the evaluation of each feature subset results in overfitting. Table 3.2-1 presents description of reviewed wrapper approaches.

Table 3.2-1: Literature Review on Wrapper Approaches for Feature (Genes) Selection

Literature	**Proposed Wrapper Gene selection approach**	**Classifier**	**Microarray Cancer Datasets (Binary Class)**	**Performance Validation Criteria**
[36]	FF-SVM	▪ SVM	▪ Colon ▪ Leukemia ▪ Lung	LOOCV
[35]	PSO-SVM	▪ SVM	▪ Colon ▪ Leukemia ▪ Lung	LOOCV
[35]	ABC-SVM	▪ SVM	▪ Colon ▪ Leukemia ▪ Lung	LOOCV
[38]	ACO-SVM	▪ NB ▪ DT ▪ kNN ▪ Rules Induction	▪ Colon ▪ CNS ▪ Leukemia ▪ Breast Cancer ▪ Ovarian	F-measure, ROC
[38]	GA-SVM	▪ SVM	▪ Colon ▪ Leukemia ▪ Lung	LOOCV

[37]	HSGAFS	▪ NN	▪ Colon ▪ Lymphoma ▪ Leukemia	10 fold Cross Validation
[39]	BPSO-SVM	▪ SVM	▪ Leukemia ▪ Prostate ▪ Colon ▪ Lung ▪ Lymphoma	LOOCV
[39]	BPSO-CGA	▪ kNN	▪ Colon ▪ Leukemia ▪ Prostate	LOOCV
[40]	HPSO-LS	▪ 1NN	▪ Colon ▪ Lymphoma ▪ Leukemia	Classification Accuracy

3.3 Hybrid Approaches for Feature Selection

Hybrid approach for feature selection is either combination of same or two or more different techniques, with an aim to combine the best traits of combined feature selection approaches. The most common hybrid combination is of filter and wrapper. In which filters are generally employed as a preprocessor for the later stage i.e. wrapper, as it statistically scores the features and pool out the informative genes using less computational resources. Features with high ranking or scores are saved and used as initial search space for wrappers, that uses the heuristic of learning algorithms to calculate the prediction accuracies of different feature candidate subsets and opt out the most accurate candidate subset. Practically, any combination of filter and wrapper can be used for constructing a hybrid but in literature variety of novel and efficient combinations of filter and wrapper has been suggested. Most of researchers have employed bio- inspired evolutionary method as wrappers like Genetic Algorithm [41-47], Particle Swarm Optimization [47 - 49], Ant Colony Optimization [50 - 52], and Artificial Bee Colony Optimization [53] and many more [54] in hybrid framework.

3.3.1 Metaheuristic Approaches for Feature Selection

Traditionally Forward and Backward search strategies with induction algorithm are used as wrapper approach for feature selection [59], but now the game has changed and metaheuristic approaches has taken the charge. Metaheuristic approaches are generally bio-inspired evolutionary approaches that follow the natural mechanism to process the data and reduce the dimensions. These approaches have played a vital role in feature selection of high dimensional datasets especially microarray cancer datasets. Few of them are: -

- Genetic Algorithm
- Particle Swarm Optimization
- Ant Colony Optimization
- Artificial Bee Colony

The basic problem associated with metaheuristic wrapper approaches is their computational cost and memory requirements. As almost all of these metaheuristics approaches involve population based parameters that require lot of memory. And to achieve optimal subset the wrapper approach has to perform subset evaluation for a number of iterations, thus induction algorithm is called again and again and this factor makes it computationally intensive [35-40]. To overcome these issues, the best approach is to preprocess the raw data using filter feature selection approaches that are computationally less expensive and play a vital role in extracting out the irrelevant and noisy features that are not adding any information to the model training. Thus search space of meta heuristics approach gets reduced and ultimately the computational and space requirements too. The summary of existing metaheuristic hybrid approaches in literature along with publisher and publishing year is presented in table 3.3-1.

Table 3.3-1: Existing Metaheuristic Hybrid Approaches

Literature	**Publisher**	**Year**	**Existing Hybrid Feature (Gene) Selection Approach**	**Filter**	**Wrapper**
[41]	**Science Direct**	2017	Information Gain and Genetic Algorithm (IG-GA)	Information Gain	Genetic Algorithm

[42]	**Science Direct**	2017	Mutual Information Maximization and Adaptive Genetic Algorithm (MIMAGA)	Mutual Information (MI)	Genetic Algorithm (GA)
[43]	**Science Direct**	2017	Laplacian Score and AI tuned Genetic Algorithm (IDGA – L/F)	Fisher Score/ Laplacian Score	Genetic Algorithm and Artificial Intelligence (IDGA)
[44]	**IEEE**	2017	Fast Correlation Based Filter and Genetic Algorithm (FCBF-GA)	Fast Correlation Based Filter	Genetic Algorithm
[46]	**IEEE**	2018	ReliefF and MIM Filters with Extended GA wrappers	ReliefF and Mutual Information Maximization	Genetic Algorithm
[44]	**IEEE**	2017	Fast Correlation Based Filter and Particle Swarm Optimization (FCBF-PSO)	Fast Correlation Based Filter	Particle Swarm Optimization
[49]	**Science Direct**	2017	Minimum Redundancy Maximum Relevance with Particle Swarm Optimization (mRMR-PSO)	Minimum Redundancy Maximum Relevance(mRMR)	Particle Swarm Optimization
[50]	**Science Direct**	2016	Fisher Criterion and Cellular Learning Algorithm with Ant Colony Optimization (CLACOFS)	Fisher Score	Ant Colony Optimization
[51]	**Springer**	2018	Mutual Information (MI) and Adaptive Stem Cell Optimization (ASCO)	Mutual Information	Ant Colony Optimization

[52]	**Springer**	2019	ReliefF with Ant Colony Optimization	ReliefF	Ant Colony Optimization
[48]	**Hindawi**	2015	Minimum Redundancy and Maximum Relevance with Artificial Bee Colony Optimization (mRMR - ABC)	Minimum Redundancy Maximum Relevance(mRMR)	Artificial Bee Colony Optimization
[53]	**Science Direct**	2017	Independent Component Analysis and Artificial Bee Colony Optimization (ICA + ABC)	Independent Component Analysis (ICA)	Artificial Bee Colony Optimization
[54]	**Taylor and Francis**	2014	Symmetrical Uncertainty and Harmonic Search Algorithm (SU-HSA)	Symmetrical Uncertainty	Harmonic Search Algorithm
[55]	**IEEE**	2016	Random Forest Ranking and Binary Black Hole Algorithm (RFR-BBHA)	Random Forest Ranking	Black Hole Algorithm
[56]	**Science Direct**	2017	Logarithm Transformation-Grasshopper Optimization Algorithm (Log-GOA)	Logarithm Transformation	Grasshopper Optimization Algorithm (Log-GOA)

3.3.1.1 Genetic Algorithm (GA) Based Hybrid Approaches

Genetic Algorithm is a bio-inspired metaheuristic approach inspired by natural evolution process. The basic idea behind genetic algorithm is to look for the fittest individual (best solution) from the search space over the generations. And the selected pair of fittest

individuals at each generation is used as parents for the next generation. The optimization of individuals is carried out using crossover and mutation operators. This process is carried out till an optimum solution is obtained or termination point has achieved. In machine learning, one of the application of genetic algorithm is feature selection. Feature selection is a combinatorial problem and genetic algorithm works perfectly for combinatorial problems that why in literature, it's been widely used either own its own as a wrapper or in combination with filters in a hybrid framework to perform feature selection specifically for the feature selection of microarray cancer datasets.

Hanaa Salem et al has combined Information gain (IG) [41], a univariate filter with Standard Genetic algorithm (SGA), a wrapper, to construct a hybrid for feature selection of microarray cancer dataset. And genetic programming (GP) is used for cancer classification with 10 folds cross validation.

Huijuan et al [42] has proposed a hybrid approach (MIMAGA) composed of Mutual Information Maximization (MIM) filter and Adaptive Genetic Algorithm based wrapper. In earlier stage, MIM is used repeatedly for genetic filtering with an aim to filter the genes with maximum information dependencies to other genes among the same class. 300 such genes are selected from this stage and then passed to the later stage AGA where optimization is carried out and classification accuracy of Extensive Learning Machine (ELM) is used to score the fitness of each individual. MIMAGA is tested on 6 Microarray datasets and for each dataset it yields reasonable number of genes and accuracy greater than 80%.

M. Dashtban et al [43] has proposed an intelligent dynamic genetic algorithm (IDGA), it's a hybrid approach in which Fisher score or Laplacian score filter is used for feature ranking and initial dimensional reduction. Genetic algorithm with dynamic parameterization and improved operators incorporating behaviorist psychology inspired penalizing strategy for obtaining crossover and mutation probability is proposed in second stage. Effectiveness of proposed algorithm IDGA is tested on 5 microarray cancer gene expression datasets. Moreover, the researcher has used and analyzed the performance of various classifiers SVM, kNN, Naïve Bayesian in combination with proposed algorithm and one of filter ranking technique. And Fisher score – IDGA - SVM combination has performed the best of all.

Table 3.3-2 presents Objective, Features, Classification Accuracy Range, Datasets, Parameter tuning of Filter and wrapper, and other traits of GA wrapper based hybrid approach.

Table 3.3-2: Summary of traits of GA wrapper based Hybrid approaches

Feature Selection Methods	**Objective**	**Feature (Genes) Count Range**	**Classification Accuracy Range**	**Datasets**	**Parameter Tuning of Filters and Genetic Algorithm Operators**	**Filter**	**Resultant Features shared**
IG-GA [41]	Improvement in Classification Accuracy	3-60	84.48% - 97.06%	Binary Class	Variable	Single	No
MIMAGA [42]	Redundancy Removal and High Dimension Reduction, Optimization of Parameters	3-216	80.4% - 97.62%	Binary Class and Multi-class Datasets	Adaptive	Single	No
IDGA_L/ IDGA_F [43]	AI Based Parameter Tuning	8-31	89.3%-98.6%	Binary Class and Multi-class Dataset	Adaptive	Choice of Filter (One at a time)	Yes
SCC-GA [44]	Improvement in Classificatio	NAN	85.24%-89.02%	Binary Class and Multi-	Partial	Single	No

	n Accuracy			class Dataset			
mRMR-GA [45]	Feature Selection and Improvement in Classification Accuracy	NAN	85.48% - 98.61	Binary Class and Multiclass Dataset	Fixed	Single (Multi-variate)	No
Extended GA [46]	Improvement in Classification Accuracy	7-124	95.2 – 97.4	Binary Class Datasets	Self-Adjusting	Fused	No

3.3.1.2 Particle Swarm Optimization (PSO) Based Hybrid Approaches

Particle swarm optimization is a global stochastic optimization technique that is inspired by the natural flocking behavior of birds and fish schooling. In this technique, each particle looks for the best position or adjust itself to the position of the best known fit particle in the search space, and ultimately all particles converge at the global optimum point in the whole search space representing a global optimum solution. We have also reviewed the hybrid techniques based on particle swarm optimization [48 - 50]

Nur et al [49] paper presents a hybrid approach in which feature ranking is performed using mRMR filter while for optimization of feature subset three types of meta-heuristics techniques i.e. Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) and Cuckoo Search (CS) are used. Experiment is carried out on Microarray cancer datasets. And to evaluate the performance of selected genes, kNN and SVM with 10-fold cross validation are employed.

Moradi et al [50] has presented a hybrid approach HPSO- LS with an objective to achieve high classification accuracy. It has improved the global search mechanism of PSO by introducing Local search algorithm that protects PSO from getting trapped in local optimum.

3.3.1.3 Ant Colony Optimization (ACO) Based Hybrid Approaches

It's a metaheuristic approach that follows the foraging behavior of social ant. The ant follows a path to the food source and each path followed is represented by a pheromone's density. The density of pheromone represents the quality of solution, so higher the density, higher will be the quality of solution to the problem of interest. The path with higher pheromone's density is followed by all other ants.

Fatemah at el [50] has proposed a combination of Fisher score, a filter, and Ant Colony Optimization and Cellular Automata (ACO-CA) based wrapper. CA is used to model the interactions among genes in complex high dimensional datasets, while Ant Colony Optimization is used for studying the structure and rule generated by CA. Basic aim of the proposed algorithm is to extract the minimal subset of informative genes. SVM, Naïve Bayesian and kNN classifiers are used to test the potential of minimal candidate subset to classify the diseases. Experiments have been carried out on 4 Microarray Cancer Datasets including 2 class Prostate dataset, and 2, 3, 4 classes Leukemia Dataset.

S. Arul Antran Vijay et al [51] has proposed a fuzzy classification system: Hybrid Stem Cell (HSC) based on Ant Colony Optimization (ACO) and Adaptive Stem Cell Optimization for Microarray gene expression dataset analysis. Mutual Information (MI), a filter approach, is used in initial stage as a preprocessor to extract informative genes. The proposed method has been tested on five microarray datasets. It performed comparatively better than Hybrid Colony Algorithm (HCA).

3.3.1.4 Artificial Bee Colony (ABC) Based Hybrid Approaches

It's a similar approach to particle swarm optimization but it is inspired by the foraging behavior of honey bees. In ABC, there are three kinds of bees; employed, onlooker and scout bee, every bee performs a particular operation. The bee responsible for the search of food source is employed bee, the one that decides and pick the fittest food source is onlooker bee, the one that scans the new food source is a scout bee, three of them work in correspondence to get to the ideal food source i.e. the globally optimum solution.

Hala et al [48] has suggested minimum redundancy maximum relevance (mRMR) and Artificial Bee Colony optimization (ABC) based hybrid approach for analysis of Microarray

gene expression datasets. This paper used bio-inspired evolutionary algorithm Artificial Bee Colony for the first time for analysis of microarray dataset in combination with SVM for classification. The combination was extensively tested on 6 microarray datasets, it performed well. And Hala et al research also attracted other researchers to try Artificial Bee Colony for optimal selection of genes in combination with other filter approaches and classifiers.

Rabia Aziz et al [52] has also proposed an Artificial Bee Colony (ABC) bio-inspired wrapper based hybrid approach for microarray dataset analysis but in combination with Independent Component Analysis (ICA) filter and Experiment has been carried out on 6 cancer gene expression datasets. Naïve Bayesian Classifier is used to test the prediction accuracy of features selected by ICA + ABC combination. It performed better than other ABC based hybrid approaches and improved performance of Naïve Bayesian Classifier.

3.3.1.5 Other Metaheuristic Approaches

Few researchers have also tried combination of these conventional metaheuristic techniques to multiply the benefits of combined technique like GBC [57]. Earlier we have presented two separate hybrid approaches mRMR-ABC and mRMR-GA from literature, but this paper Hala et al. [57] has presented a novel hybrid approach by combining the Artificial Bee Colony and Genetic Algorithm I.e. Genetic Bee Colony optimization (GBC). The paper has improved the exploitation and exploration capabilities of basic ABC method by introducing the crossover and mutation operators of GA in it, sprout bee count is also made 2 to enhance the movement, and to enhance the computational effectiveness of this wrapper, the researcher has combined it with Minimum Redundancy Maximum Relevance Filter approach for initial preprocessing of datasets. In First Comparison of this novel hybrid has proved it superiority over other GA and PSO based hybrid approaches.

Some new metaheuristic techniques including Cuckoo search, Harmonic Search, Bat Algorithm have also been introduced by researchers for feature selection. This paper Elnaz Pashaei et al [55] has proposed Random Forest feature ranking and Binary black hole algorithm based novel hybrid approach. Bagging with 10-fold cross validation is used as a classifier. Experimentation on 7 benchmark gene expression dataset has been carried out to test the effective ness of selected genes.

Another filter-wrapper hybrid technique is presented in literature by Lin et al [70] based on Fisher Markov filter and Multi objective binary biogeography based gene selection (MOBBBO) with SVM classifier to generate set of informative genes for classification of 10 microarray cancer datasets.

Salam Salameh Shreem et al [54] has proposed a hybrid of symmetric uncertainty (SU) and a meta-heuristic search approach Harmony search algorithm(HSA), collectively termed as SU-HSA. HSA is combined with two classifiers NB and IB1. And performance of proposed feature selection hybrid is tested in combination with each classifier and also compared with other hybrid approaches. The experiment was carried out on 5 Microarray cancer datasets.

This paper [58] proposed a mRMR, a filter and Flower Pollination Algorithm, a wrapper based hybrid approach. FPA mimics the natural process of flower pollination for gene selection. 50 top ranked genes are selected by mRMR and then passed to the later FPA wrapper stage for optimization, the selected gene subset is evaluated by SVM classifier with 10-fold cross validation. The researcher has also compared the results with mRMR-GA and both have almost performed equally good.

Table 3.3-3: Description of Hybrid Approaches for Feature (Genes) Selection

Literature	Proposed Hybrid Gene Selection Approach	Classifier	Microarray Cancer Datasets	Performance Validation Criteria
[41]	**Information Gain and Genetic Algorithm (IG-GA)**	▪ Genetic Programming (GP)	▪ Leukemia ▪ Colon ▪ CNS ▪ Lung Ontario ▪ Lung Michigan ▪ DLBCL ▪ Prostate	10 Fold Cross Validation
[42]	**Mutual Information Maximization and Adaptive Genetic Algorithm (MIMAGA)**	▪ ELM ▪ RELM ▪ BP ▪ SVM	▪ Leukemia ▪ Colon ▪ Prostate ▪ Breast	Classification Accuracy
[43]	**Fisher Score and AI tuned Genetic Algorithm (IDGA – F)**	▪ SVM ▪ kNN ▪ NB	▪ Breast ▪ DLBCL ▪ Leukemia	LOOCV
[43]	**Laplacian Score and AI tuned Genetic Algorithm (IDGA – L)**	▪ SVM ▪ kNN ▪ NB	▪ Breast ▪ DLBCL ▪ Leukemia	LOOCV
[44]	**Spearman's Correlation Coefficient**	▪ SVM ▪ kNN ▪ NB ▪ DT	▪ Colon ▪ SRBCT	Classification Accuracy
[45]	**mRMR-GA**	▪ SVM	▪ Lymphoma ▪ Colon ▪ Lungs	LOOCV
[47]	**Fast Correlation Based Filter and Genetic Algorithm (FCBF-GA)**	▪ SVM	▪ Colon ▪ DLBCL	LOOCV
[51]	**Fisher Criterion and Cellular Learning Algorithm with Ant Colony Optimization (CLACOFS)**	▪ NB ▪ kNN ▪ SVM	▪ Leukemia ▪ Prostate	ROC curve
[52]	**Mutual Information (MI) and Adaptive Stem Cell Optimization (ASCO)**	▪ Fuzzy Classifier	▪ ISR ▪ T2D ▪ Colon ▪ Leukemia ▪ Prostate	Classification Accuracy of Rules Generated
[47]	**Fast Correlation Based Filter and Particle**	▪ SVM	▪ Colon ▪ DLBCL	LOOCV

	Swarm Optimization (FCBF-PSO)			
[53]	**Independent Component Analysis and Artificial Bee Colony Optimization (ICA + ABC)**	▪ Naïve Bayesian (NB)	▪ Colon ▪ Leukemia ▪ Prostate ▪ Glioma ▪ Lung	LOOCV
[48]	**Minimum Redundancy and Maximum Relevance with Artificial Bee Colony Optimization (mRMR - ABC)**	▪ SVM	▪ Colon ▪ Leukemia ▪ Lung	LOOCV
[54]	**Symmetrical Uncertainty and Harmonic Search Algorithm (SU-HSA)**	▪ Instance Based Classifier (IB1) ▪ NB	▪ Leukemia ▪ Colon ▪ CNS ▪ Breast ▪ Ovarian	10 Fold Cross Validation
[55]	**Random Forest Ranking and Binary Black Hole Algorithm (RFR-BBHA)**	▪ Bagging	▪ Colon ▪ CNS	10 Fold Cross Validation

3.4 Analysis

From Literature, we can conclude that every feature selection approach has its own pros and cons. Filter based feature selection has more generalization properties as compared to other approaches but they lack the capabilities to reduce the dimensions in case of high dimensional datasets and thus do not generate the good prediction accuracies [28-35]. To overcome the drawbacks associated with filter approaches, wrapper and hybrid approaches are proposed that involve the heuristics of classifiers to evaluate the performance of selected features. Wrapper methods [35-40] are better alternative to filters for supervised learning problems being efficient in performance but are computationally expensive, hence require plenty of computational resources for high dimensional datasets. Moreover, wrapper models are prone to overfitting, calling classifier again and again for the evaluation of each feature subset results in overfitting.

The hybrid methods seem the most promising gene selection methods for microarray datasets in terms of classification accuracies, no. of selected genes and computational costs. Hybrid Feature selection approach encompass the strength of both filters and wrappers. Thus

involve multiple evaluation criteria at both stages. There are so many algorithms and thus this area of research and room for trying new combination never get narrowed.

We have done a thorough literature review on hybrid approaches based on GA, PSO, ABC, ACO and few others [41-58]. These combinations have ensured correct diagnosis with few no. of features and have lower risk of overfitting.

GA is considered as old and most promising feature selection metaheuristic approach [59] but with reference to table 3.3-2, we have seen it's hybrid version has not yet achieved exceptional classification accuracies and features count. Moreover, the gaps we have analysed in literature [42-46] that almost all GA hybrid approaches combine only one filter approach with wrapper as a preprocessing technique and this the pattern we have seen in almost all GA based wrapper approaches. This may result in biased output as we are just considering the heuristic of one filter approach. Every Feature selection technique has its own way of evaluation and ranking criteria. The output can be more effective if we consider the heuristics and evaluation criteria of more than one filter as we have done in our proposed approach. Secondly restricting No. of feature to some k value may cause potential feature to lose. And this k value is usually based on user's choice.

These gaps give birth to our problem statement i.e. improvement in the performance of Traditional Genetic Algorithm Wrapper. Few researches have covered the gap of performance by combining multiple metaheuristic approaches as done by Hala et.al [57], another approach [46] has extended by creating a hybrid of filters and ensemble but this approach becomes expensive in terms of time complexity and execution time, as multiple wrappers operate in ensemble framework. So we have decided to bring improvement in traditional Genetic Algorithm approach by combining multiple univariate information theory based filters as a preprocessing step of wrapper that are relatively less computationally expensive than wrappers. We have also ensured in this stage not to compromise over relevancy. And one other problem that is associated with filter is redundancy, we have also overcome it by using Pearson correlation statistics. Moreover, we have seen generally SVM is used and induction algorithm for GA wrapper, we have brought novelty by introducing Random forest classifiers as induction algorithm. In later section we have covered comparison of GA-SVM and GA-RF approach to justify the effectiveness of our choice of induction algorithm.

CHAPTER 4: PROPOSED FRAMEWORK
MULTIPLE FILTERS AND GA WRAPPER WITH RF (MF-GARF)

In this section, we present a hybrid approach for feature selection of microarray cancer dataset that preserves the advantages of filter and wrapper methods while mitigating their drawbacks. A schema of proposed framework is given in Figure 3.4-1.

In the first stage, we have used a set of three information based filter techniques Information Gain, Gain Ratio and Gini Index, each of these filter technique score each feature statistically without any learning algorithm and selects the top-scoring features filtered by each filter method, meeting a specific threshold criterion. A feature set is then created by taking union of features opted by each filtering technique. All three filters rank feature based on information they add to the class label, so directly ensure the relevancy of selected feature to the class label. Another filtering technique Pearson Correlation is used that removes the redundancy from the selected features. Thus turning a high dimensional dataset into a small amount of feature pool, serving as a reduced search space for an optimal wrapper approach Genetic Algorithm that incorporates the Random Forest to evaluate the fitness of each selected feature subset. We have used set of univariate filters that score each feature individually thus do not consider the relationship among feature, the subset of feature may bring more information to the leaning model instead of an isolated feature but this may induce redundancy.

Feature subset selected by filters can be still large and it's not tuned to any classifier that's why we introduced a second stage where a wrapper is used to reduce the dimensionality of the feature subset. Motivation of this hybrid approach is to involve both important aspects of feature selection i.e. relevancy and redundancy analysis of features. And bring forth an optimal subset of features. GA Wrapper Approaches are computationally expensive for high dimensional datasets, that why in the initial stage we have used filters that serve as a pre-processing step for wrapper that reduces its search space and improves it performance. Thus this approach of feature selection along with high dimensionality reduction provide a time and space complexity improvement.

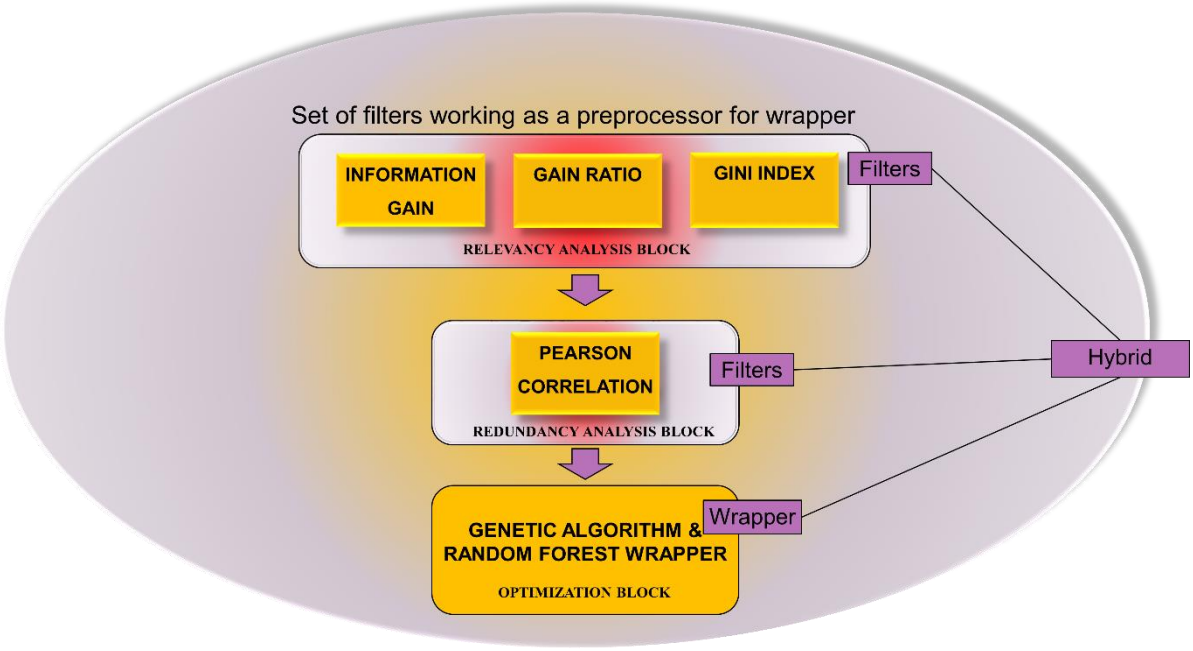

Figure 3.4-1: Schema of Proposed Framework MF-GARF

4.1 Relevancy Analysis Block

In this block, relevancy of each feature (gene) with the target class is calculated using information FEF based univariate filters Information gain, Gini index, and Gain Ratio.

4.1.1 Information Gain

Information gain [9] is one of the most preferred feature ranking filter that measure the relevancy of each feature and helps to make decision either the feature should be chosen or not. Information gain is a symmetrical measure of mutual dependence between two variables. It captures information regarding one random variable, through other random variable. It is one of the variants used by decision tree in machine learning to capture the importance of features. Information gain is based on entropy, which is measure of randomness in the information being processed, where there is a minimum entropy there is a maximum information gain. For each feature information gain value is calculated. Greater value of Information gain depicts relevancy of feature to the target class. A threshold criterion is adjusted to make a choice of features to be kept, a feature with information gain value above or equal to threshold value are kept while others are discarded.

$$Entropy = -\sum_{i=1}^{c} (p_i \log_2(p_i)) \quad (4.1)$$

$$Information\ Gain\ (D_P, f) = Entropy(D_P) - \frac{N_{left}}{N} Entropy(D_{left}) - \frac{N_{right}}{N} I(D_{right}) \quad (4.2)$$

4.1.2 Gain Ratio

Gain Ratio [10], a term coined by Ross Quinlan, is an improved version of Information Gain. It scores the features in a similar way as the Information gain, calculates the information regarding one random variable *(x)*, through other random variable. But Gain ratio involves intrinsic information in order to give overall score to each feature. Information gain favours multi-valued features. The approach of gain ratio is to amplify the information gain while limiting the number of its values. The equation for gain ratio is as follow: -

$$Gain\ Ratio(x) = \frac{\text{Information Gain (x)}}{\text{Intrinsic Value (x)}} \quad (4.3)$$

4.1.3 Gini Index

Gini Index [11] is a term coined by Corrado Gini, an Italian statistician in 1912. It measures the impurity and uncertainty among the values of the features. Greater the value of an index, more the data will be distributed. This measure of impurity is used by Classification and Regression Tree (CART). This technique is considered to be more fast as compared to other filtering technique. We have used Gini index to evaluate the goodness of each feature. The basic purpose behind using this approach is to select those features having minimum Gini index value and thus bringing in maximum purity improvement. The feature with minimum Gini index value is considered as a splitting point in CART as it brings maximum purity. The equation for the Gini Index is as follow: -

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2 \quad (4.4)$$

For each filter approach following steps are followed:

1. Microarray cancer dataset is retrieved.
2. Missing values are removed.

3. Dataset is passed to filter that evaluates the feature according to its evaluation criteria, and evaluation score is assigned to each feature.
4. Evaluation Scores are normalized using formula.
5. Now feature scores are compared with threshold criterion i.e. *th >= 0.5*. Feature satisfying the threshold criteria are kept while others are discarded.

At final step, all the feature sets are merged into a unified set that serves as an input for redundancy analysis block.

4.2 Redundancy Analysis Block

The above used filter approaches are univariate, they evaluate and score features independently without considering feature interaction, so there are chances of presence of highly correlated features that induce redundancy in data and add no additional information to the model, hence these features are undesirable. For that we have incorporated Pearson Correlation coefficient measure to calculate the correlation among features.

4.2.1 Pearson Correlation

Pearson Correlation [12] has been used to overcome the issue of redundancy among the features. The aim is to come up with a feature subset incorporating features that are highly correlated with the target class but have low correlation with each other. So using Pearson correlation, that is one of the most helpful statistically measure for figuring out the strength and relationship among variable, correlated features are removed as they do not add any additional information to the learning model. Individually they might have some presence but there exist other features similar in behavior, having same impact on prediction, thus resulting in redundancy. Removing correlated features saves space and time of calculation of complex algorithms. Moreover, it also makes processes easier to design, analyze, understand and comprehend. The equation for Pearson Correlation (r) is as follow: -

$$Pearson\ Correlation\ (r) = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (4.5)$$

Features with high correlation value i.e. greater and equal to 0.85 with other features are removed from the set D, and new set D' is obtained.

The set D' serves as search space of Genetic algorithm wrapper in optimization block.

4.3 Optimization Block

The optimization block is comprised of genetic algorithm wrapper with random forest as an induction algorithm to bring forth an optimal set of features (genes).

4.3.1 Genetic Algorithm

Genetic Algorithm based on Charles Darwin theory of natural evolution, follows the pattern of natural selection to select the individuals with maximum fitness score to reproduce offspring for the next generation. Here from fitness score of individual we mean those individual that have more power to withstand the environmental changes. Genetic algorithm (GA) [41-48] being one of the most promising and efficient optimization technique has been used in combination with many wrapper and hybrid methods for feature selection and classification of high dimensional dataset, especially microarray datasets, for last two decades.

Genetic Algorithm [41-48] performs a refined feature selection from a pool of highly informative features. Genetic Algorithm is a global search technique that improve the quality of selected feature by finding an optimal feature subset. It looks into the search space for the fittest feature subset that produces the best classification accuracy. The initial population, fitness function, selection, crossover and mutation operator are the five main components of the genetic algorithm. The population of Genetic Algorithm is comprised of chromosomes, and each chromosome in population corresponds to a solution to the optimization problem. Each chromosome is incorporated as a binary sequence (0's and 1's). The length of chromosome corresponds to the number of features in dataset, the presence of feature is represented by 1, while 0 indicates the absence of feature. And role of each chromosome for the next generation is determined by the fitness value it acquires.

In our proposed hybrid approach, the fitness is measured as a function of the accuracy of the Random Forest classifier with which GA is wrapped in 10-fold cross validation setup. For fitness evaluation of each feature subset Random Forest, an ensemble model, is used. It has been discussed in later section.

The mutation probability (Pm) is used to avoid the trapping of Genetic Algorithm in local minima. The value of mutation probability is typically kept low [60], i.e. lies within a range

0.001 to 0.01. High mutation probability converts GA into a random search. For this study, we have kept 0.5 as a mutation probability.

The crossover probability (Pc) is suggested with in a range 0.5- 0.9 [60], we have opted 0.5, and single point crossover as a crossover way.

Tournament selection is chosen as a selection scheme with a value 0.25, keeping tournament size low helps to keep up the diversity within the population, otherwise high tournament size increases pressure and results in decrease in population diversity.

4.3.2 Fitness Function

In GA [60], fitness function is defined to measure the fitness of each individual chromosome so as to determine which will reproduce and survive into the next generation. Thus, given a particular chromosome, the fitness function returns a single numerical score, 'fitness', which is proportional to the 'ability' of the individual that the chromosome represents. The 'fitness' score assigned to each individual in the population depends on how well that individual solves a specific problem.

In our proposed approach we have used random forest classifier as an induction algorithm for GA wrapper to evaluate the fitness of each chromosome x, the fitness of x is measured as a function of the accuracy of the Random Forest classifier with which GA is wrapped in 10-fold cross validation setup. Random Forest [61] operates as an ensemble approach it considers "the wisdom of crowd". It builds multiple uncorrelated decision trees, each decision tree individually predicts the class, final prediction of class is based on the majority vote. Generally, researchers have used SVM as induction algorithm but considering the positive traits of Random Forest we have preferred it over SVM. We have also shown a comparison of GA-RF and GA-SVM in later section. To overcome the issue of time complexity we have used Gini index (*Equation 4.5*) as a splitting criteria, that takes comparatively less time than other two splitting criteria Information Gain and Gain Ratio.

$$Fitness\ of\ each\ chromosome\ (x) = 10foldsCV(classifier_{accuracy)} \quad (4.6)$$

4.4 Classifiers

For evaluation of selected feature subset by proposed MFGARF following classifiers are used in 10-fold cross validation setup.

4.4.1 Random Forest (RF)

Random Forest [61] operates as an ensemble approach it uses decision tree as a base model. The best thing about Random forest is that it considers "the wisdom of crowd". It builds multiple uncorrelated decision trees, each decision tree individually predicts the class, final prediction of class is based on the majority vote. In our proposed approach we have used random forest classifier as an induction algorithm for GA wrapper to evaluate the fitness of each chromosome. For that random forest has used Gini index as a splitting criteria, as it does not involve logarithm thus making random forest computationally inexpensive. In later stage, Random Forest is also used for the evaluation of robustness of final optimal feature subset.

4.4.2 K Nearest Neighbor (kNN)

K Nearest Neighbor [62] is a widely used Machine Learning non-parametric technique. It performs the classification based on similarity measures, i.e. measures the distance of new sample cases from the training samples. Here we have used it to evaluate the classification performance of our selected optimal features (genes).

4.4.3 Naïve Bayes (NB)

Naive Bayes [63] is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network. Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection.

4.4.4 Support Vector Machine (SVM)

SVM [64] is also a widely preferred machine learning algorithm that work with an objective to look for a hyperplane that distinctly classify the data points. The hyperplane with

maximum margin is chosen having maximum distance from both classes data point to later classify the new coming data points with confidence. We have used SVM for classification of our selected optimal feature subset in 10-fold Cross Validation. SVM performs best with binary classes.

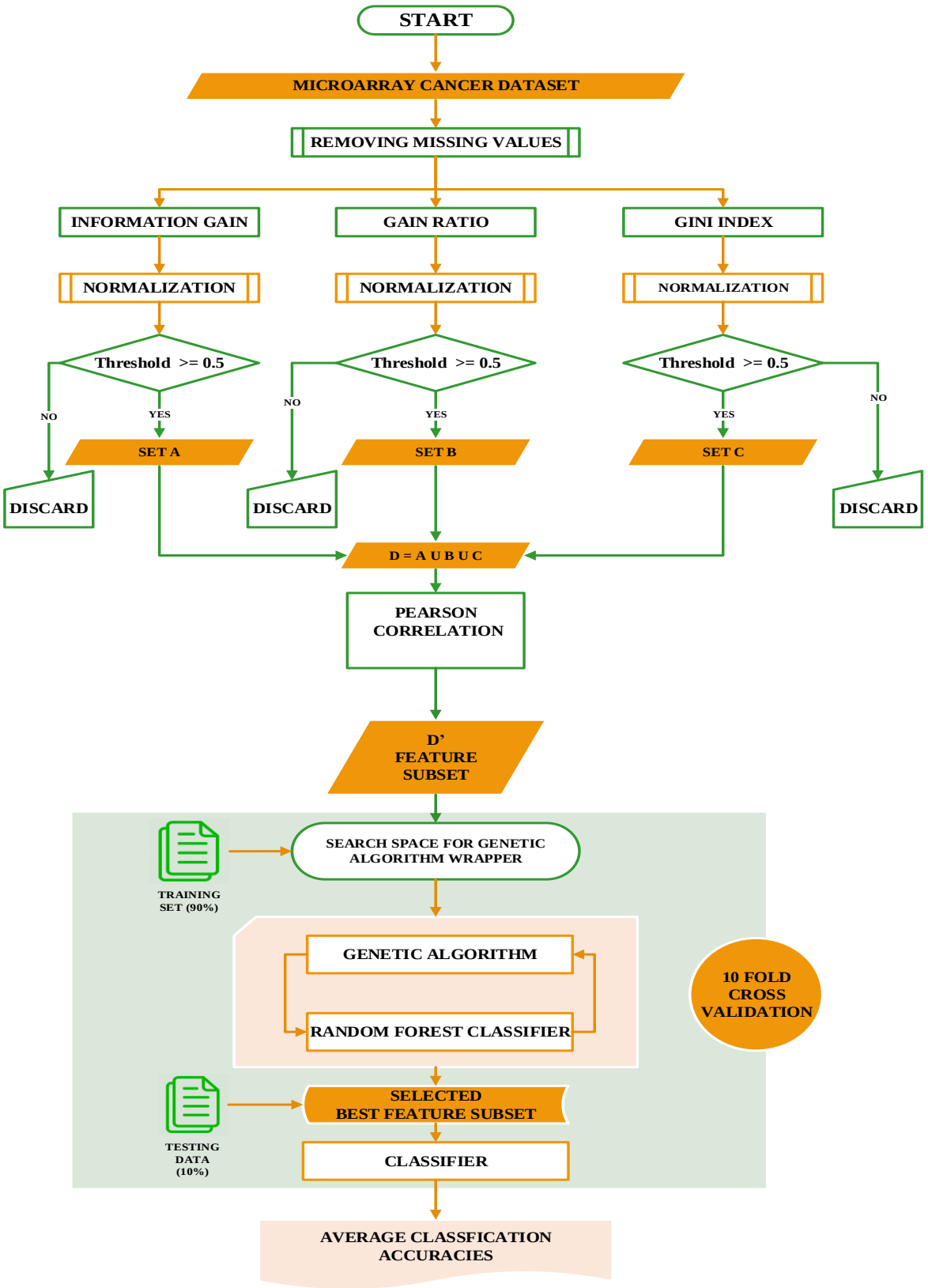

Figure 4.4-1: Flowchart of proposed MFGARF

CHAPTER 5: EXPERIMENTAL SETUP AND RESULTS

5.1 Microarray Datasets

DNA Microarray technology is a valuable advancement in medical field that facilitates medical specialists in monitoring and profiling gene expressions of an organisms. With the help of this technology, biologist can profile thousands of gene expressions in a single experiment. In this section, we have discussed the microarray gene expression datasets that have been used to test the effectiveness of our proposed hybrid approach. These datasets are publically available and have been used by many studies [27-57] for the purpose of gene selection task and classification of diagnostic classes.

The selected Microarray Cancer datasets contain 5 binary class datasets Colon [63], Prostate [64], Leukemia [63], Ovarian [63] and Central Nervous System (CNS) [63] while two multiclass datasets small round blue cell tumors (SRBCT) [63] and Lymphoma [63]. These Cancer datasets often serves as a benchmark for microarray analysis methods. The table gives a description of the chosen microarray datasets covering features count, no. of instances and imbalance ratio among classes.

Table 5.1-1: Description of Microarray Cancer Datasets

	Datasets	**Features**	**Instances**	**Classes**	**Instances Distribution**	**Imbalance Ratio**
BINARY CLASS DATASETS	COLON [65]	2000	62	Normal/Tumor	Normal: 20 Tumor: 40	1.82
	PROSTATE [66]	12533	102	Normal/Tumor	Normal: 50 Tumor: 52	1.04
	LEUKAEMIA [65]	7129	72	AML/ALL	AML: 25 ALL: 47	1.88
	OVARIAN[65]	15153	253	Normal/Cancer	Normal: 91 Cancer: 162	1.78

	CNS[65]	7129	60	1/0	1: 39 0: 21	1.85
MULTI CLASS DATSETS	SRBCT[67]	2308	83	1/2/3/4	1: 29 2: 11 3: 18 4: 25	1.9
	LYMPHOMA [65]	4026	66	DLBCL/FL/CLL	DLBCL: 46 FL: 9 CLL: 11	2.6

5.2 Experimental Setup

The framework is designed in python on Jupiter Notebook version 5.7.4. pandas, numpy, DEAP, scikitLearn libraries in python are used for implementation. We have also used implementations of Rapid Miner for experimentation. And all the experiments are carried out on Desktop-ISF1EID having Intel (R) Core (TM)-8700k CPU with 3.70 GHz processing speed, 6 Cores, 12 internal processors and 16GB RAM. Moreover, datasets used for evaluation of proposed framework, the parameter tuning of wrapper, classifier, and experimental results are discussed in this section.

5.3 Parameter Tuning

For experimentation, few parameters are tuned. In relevancy analysis block, the threshold for maintaining relevancy is set to 0.5 for all three filters. Different value of threshold for all three filters are adjusted and tested, the one that performed the best in terms of classification accuracy is considered as a final filtration criterion. In redundancy analysis block, Correlation value is set to 0.85 to extract the highly correlated features having same or above correlation value with other features. The parameters of Genetic algorithm i.e. population is set to 50, number of generations is set to 30 to iterate the process 30 times to get the most optimal subset of features. The values of probability of crossover (Pc) is set to 0.5 while probability of mutation (Pm) is assigned 0.01 value.

The parameters of random forest classifier number of trees, maximum depth and splitting criterion are set to 25, 25 and Gini Index respectively. In order to tune the parameters of rapid miner we have utilized the rapid miner analytics that suggest the values based on user's choice. Advantage of Random forest over other classifier is that it is not sensitive to parameter tuning that's why we have considered the suggested values by rapid miner. The figures 5.3-1, 5.3-2 below show the analytics provided by rapid miner.

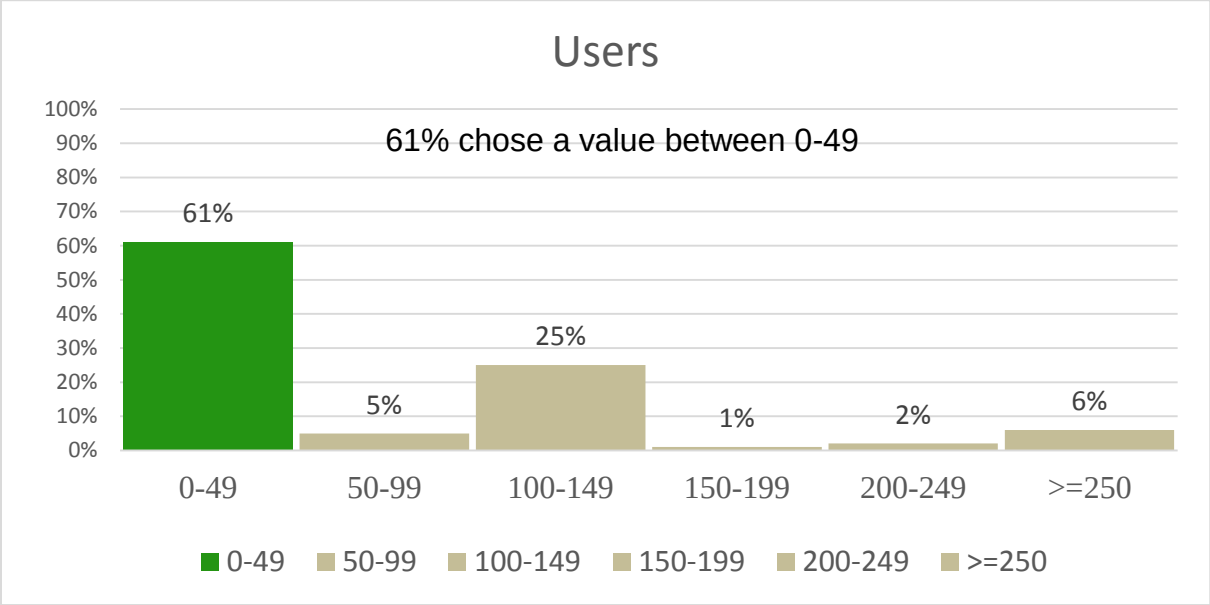

Figure 5.3-1: Parameter Tuning of Random Forest: Number of Trees

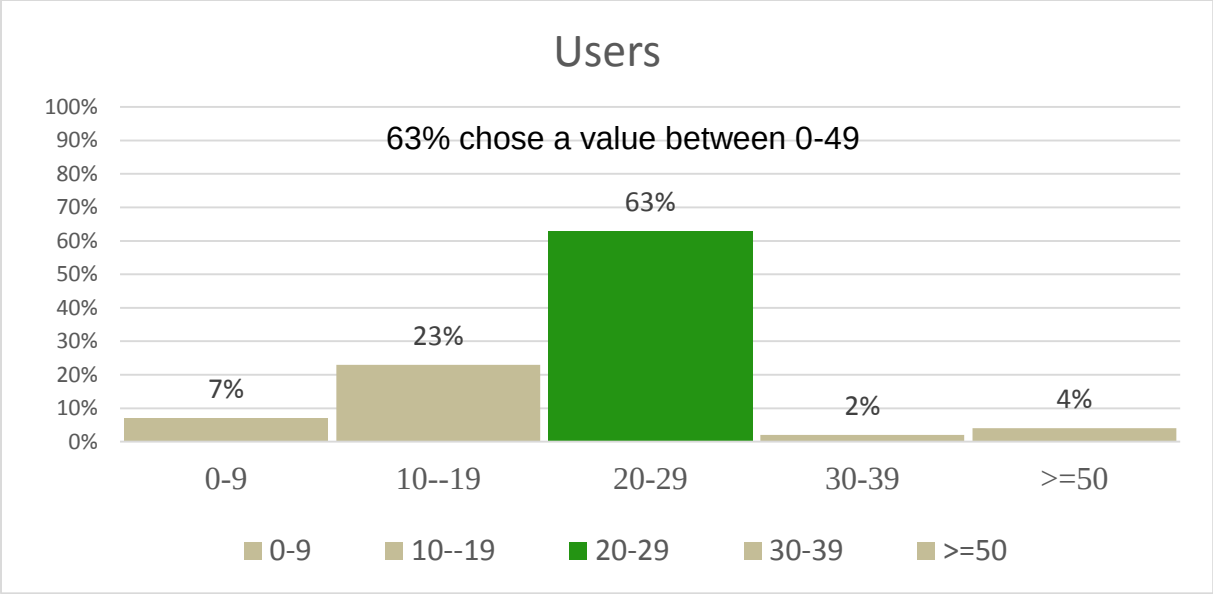

Figure 5.3-2: Parameter Tuning of Random Forest: Depth of Tree

5.3.1 Threshold Value Adjustment of Filters

The threshold value adjustment is an important aspect for filter method in order to make a decision what to keep and what to discard. We have tried each of our filter methods Information Gain, Gain Ratio and Gini Index with different threshold values and evaluated the relevancy of filtered features to predictive class by their classification performance. For that purpose, random forest is applied in 10-fold cross validation setup. The threshold value that produced the feature set with the best accuracy is considered for each filter method. In our setup, 0.5 is used as final threshold value. And thus, only those features are kept for later stage, having relevancy score assigned by each filtering feature evaluation function equal and greater to 0.5.

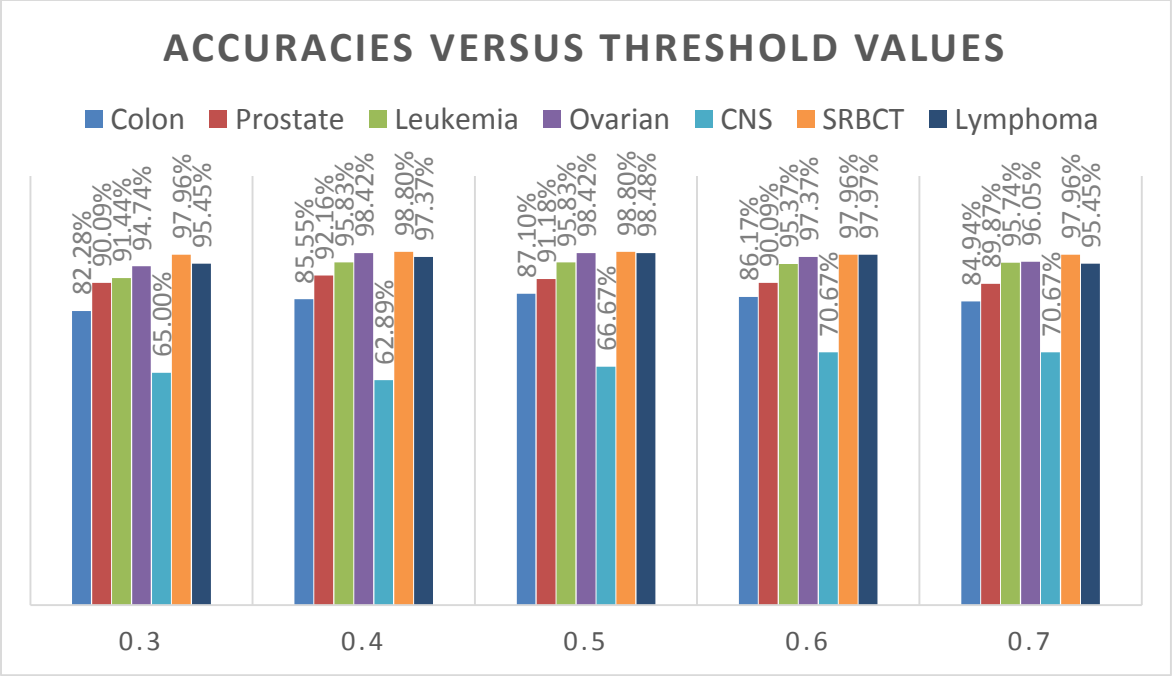

Figure 5.3-3: Accuracies versus Threshold Values

From figure 5.3-3 we can see we got better accuracies on average for all microarray cancer datasets with threshold value 0.5, so we decided to keep it as a default criterion for all of the three filters Information Gain, Gain ratio and Gini Index.

For redundancy block, where we used Pearson correlation statistics to measure the correlation among features with a motive to remove those feature that are relevant to predictive class though, but are highly correlated to other features. Such features being repetitive add no extra information to the predictive class, thus are of no use. So, we tried different threshold values too. After few experimentation, the value that satisfied our motive was 0.85 that for few cases as shown in later section, not only, removed redundancy but improved the classification accuracy too.

5.3.2 Effect of Parameter Tuning of Genetic Algorithm Wrapper

In order to make the best selection of parameters for the genetic algorithm we have tried it on various values for population size and number of generations suggested and used by different studies. Singh et al. [44] study has performed experiments by setting population size 20,30 and 50 with mutation probability 0.01. We have also performed experimentation with all three suggested Population sizes. Two studies [45] [65] has used Population Size 100, we tested

this value too. Here we have values 27 and 80 for Population and Generation Size respectively, we got it for Colon Dataset particularly by using Randomized Cross Validation is also tried for all other datasets too.

From the table 5.3-1 we can clearly see, parameter tuning has its impact on classification accuracy. The values for the parameters of the genetic algorithm that performed the best in our setup are as follow: Population Size = 50, Generation Size = 100, probability of mutation (Pm) = 0.01 and Crossover (Pc) = 0.5.
The table 5.3-1 gives an overview of the effects of parameter tuning on classification accuracies and number of selected genes. These classification accuracies are obtained by random forest classifier in 10-fold cross validation setup.

Table 5.3-1: Parameter Tuning of GA Wrapper

DATASETS		**P: 20 G: 30 [44]**	**P: 30 G: 40 [44]**	**P: 27 G: 80**	**P: 50 G: 100 [44]**	**P: 100 G: 20 [45]**	**P: 100 G:300 [65]**
COLON	Accuracy	88.89	94.44	94.44	**95.16**	88.89	88.89
	Features	10	10	10	**7**	7	7
PROSTATE	Accuracy	89.66	89.66	93.10	**97.06**	93.10	93.10
	Features	7	10	8	**10**	9	10
LEUKAEMIA	Accuracy	92.86	92.86	100	**100**	100	100
	Features	10	10	7	**6**	10	10
OVARIAN	Accuracy	98.00	96.00	98.00	**100**	**100**	**100**
	Features	3	3	4	**4**	**4**	**4**
CNS	Accuracy	83.33	83.33	88.33	90.00	**93.33**	83.33
	Features	9	10	8	10	**8**	9
SRBCT	Accuracy	98.80	100	100	**100**	98.80	**100**
	Features	10	10	10	**7**	7	**7**
LYMPHOMA	Accuracy	92.31	98.48	100	**100**	100	92.31
	Features	9	10	10	**6**	10	10

5.4 Validation Methods

Different studies have suggested different strategies for the purpose of validation of selected feature subset. As we have seen in literature mostly studies choose cross Validation [41] [49] [50] measure for the purpose of validation. It's a best practice so far, multiple iterations with random samples helps to avoid the chances of overfitting for the selected classifier. In this study we have employed stratified 10-fold cross validation to assess the classification performance for each classifier. Stratified cross validation is considered more appropriate approach than regular cross validation, it assures each fold make a good representation of the whole dataset by having equal mean response value.

In cross validation the dataset is divided into training and testing splits in ratio 90 and 10 respectively. And in each fold classification performance is evaluated.

5.5 Performance Evaluation Metrics

In order to evaluate the performance of the proposed framework, following metrics are considered in each fold of Cross Validation.

5.5.1 Classification Accuracy

We have measured the classification accuracy of selected optimal feature subset using Random Forest (RF), k-Nearest Neighbor (kNN), Naïve Bayes (NB) and Support Vector machine (SVM).

Accuracy can be calculated by using following the formula

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} * 100 \quad (5.1)$$

Here TN and TP represents the count of instances correctly predicted as negative and positive respectively, while FN and FP is for count of instances falsely predicted as negative and positive respectively

5.5.2 Precision

Precision can be defined as a percentage of correct predictions and it can be computed by using following formula.

$$\frac{TP}{TP+FP} * 100 \tag{5.2}$$

5.5.3 Recall

Recall is also known as True Positive Rate (TPR), it computes the percentage of positively predicted instances using following formula.

$$\frac{TP}{TP+FN} * 100 \tag{5.3}$$

5.5.4 AUC

Here we have used another metric for classification analysis of binary class datasets that is Area under curve (AUC). ROC curve is an application of AUC. Each point of the ROC curve is true positive rate against false positive rate for a specific applied threshold on the confidence of the corresponding classifier. The value of AUC corresponds to the goodness of model.

AUC is a scalar value between zero and one that summarizes the analysis of ROC. It is calculated according to 13.

$$\begin{aligned} \mathrm{AUC} &= \int_{\infty}^{-\infty} \mathrm{TPR}(T)(-\mathrm{FPR}'(T))\,\mathrm{d}T \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T)\mathbf{f}_1(T')\mathbf{f}_0(T)\,\mathrm{d}T'\,\mathrm{d}T = P(X_1 > X_0) \end{aligned} \tag{5.4}$$

where T is the threshold in which the instance X is classified as positive if X > T, and negative otherwise. Additionally, X1 is the score for a positive instance, and X0 is the score for a negative instance. TPR = TP/ (TP + FN) and FPR = FP/ (TN + TP) are the true positive and the false positive rates, respectively. An AUC value close to one indicates a better performance for the method. Unlike accuracy, AUC does not depend on the cutoff chosen by the classifier or on the

class distribution of the samples in the dataset. Hence, it is a more robust metric for performance evaluation.

5.6 Experimental Results

In this section we present the results of our experimentation performed on 7 benchmark Microarray gene expression datasets.

5.6.1 Case 1: Colon Cancer Dataset Experiment

The Colon Cancer Dataset based on cancer study by "*Alon et al., 1999"* [68] contains 2000 microarray gene expressions (features) against 60 patients (instances), among them 40 are with normal tissues and 22 are with tumor tissues.

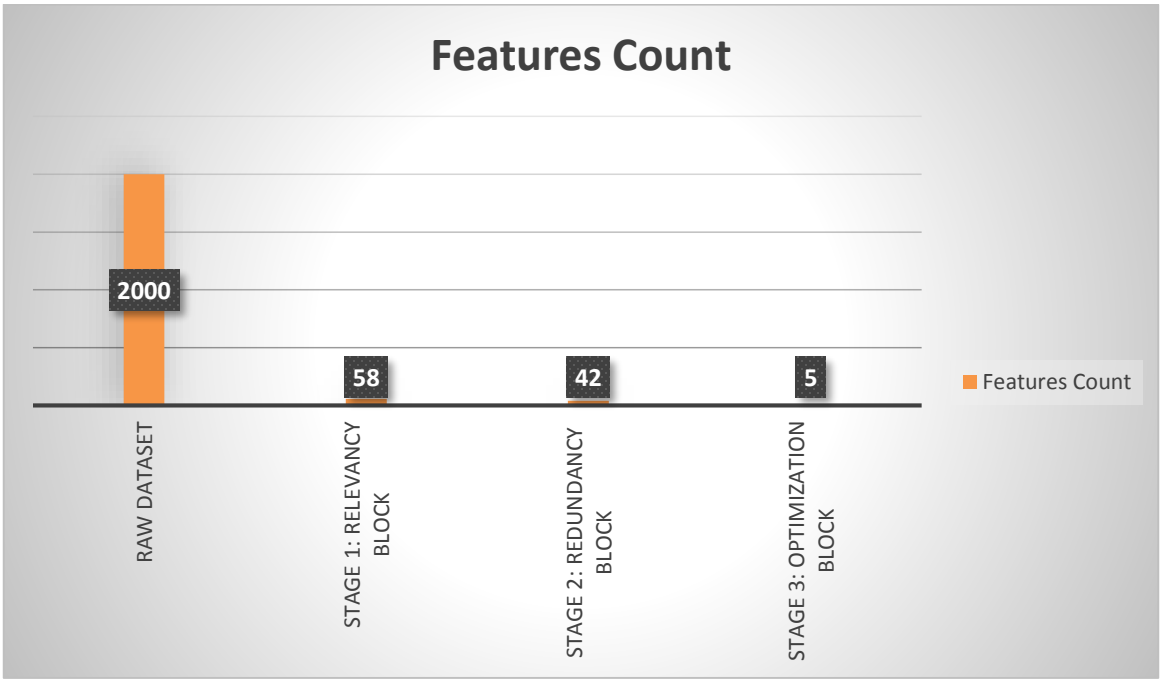

Figure 5.6-1: Colon Cancer Dataset: Feature (genes) count after each stage

The figure 5.6-1 shows the count of selected features after each stage. Raw Colon cancer dataset contains 2000 features (genes). In Relevancy Block, Information Gain Filter selected 38 Features, Gain Ratio filter selected 33 Features and Gini Index filter selected 34 Features. Union of Features by each filter meeting a specific threshold criteria resulted in a set comprising of 58 informative and relevant genes. In next stage, redundancy among features is removed and we

removed 16 genes that were redundant and not adding any new information to the target class, resulting in 42 features.

These 42 genes are passed as a reduced search space for Genetic Algorithm wrapper, where fitness of each individual is calculated by the classification accuracy of Random Forest in a 10 fold stratified Cross Validation Setup. And we got a final optimal feature subset containing 5 potential features shown in table 5.6-1.

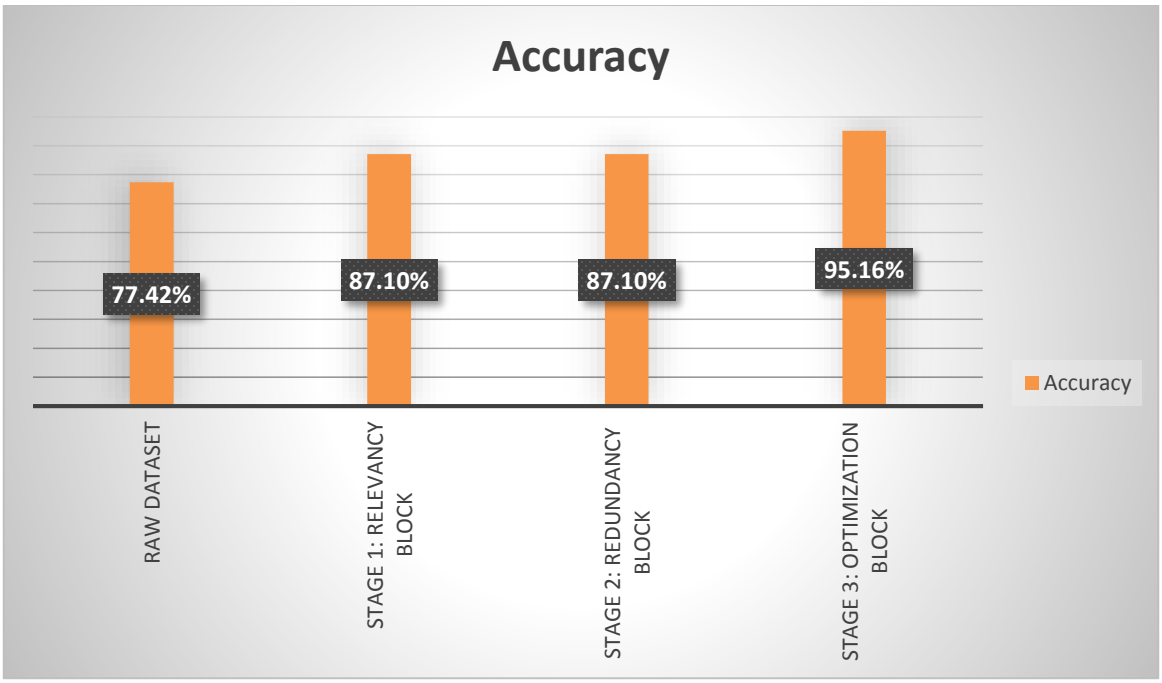

Figure 5.6-2: Colon Cancer Dataset: Classification Accuracy after each stage

Figure 5.6-2 shows classification accuracy achieved by features set obtained at the end of each stage using random forest classifier in 10 fold stratified Cross Validation setup.

To create a baseline classification accuracy, we computed accuracy with Raw Dataset without feature selection i.e. 77.94%. And to witness impact of each stage on classification performance of classifier, we have computed classification accuracies at the end of each stage. We got 87.10% accuracy for colon dataset after both stages, which clearly shows, we have removed only those features that were redundant, not contributing anymore to the predictive class. Finally, the optimal feature set has attained 95.16% classification accuracy.

Table 5.6-1: Colon Cancer Dataset: Selected optimal feature set

Selected Potential Attributes	Accuracy Achieved
{M16937, X12671, T51571, H15813, M91463}	95.16 %

Table 5.6-2 shows the accuracy, precision, recall and AUC measure for the validation split as well as for the test split. To avoid any chances of error and to assure the classification performance of selected feature subset, we have performed 10-fold Cross Validation, using different classifiers including Random Forest, SVM, Naïve Bayes and kNN. Here in the table 5.6-3 the best classification accuracy is shared which we have got with random forest Classifier.

Table 5.6-2: Colon Cancer Dataset: Confusion Matrix

VALIDATION SPLIT (CV 10 FOLDS)				**TEST SPLIT (CV 10 FOLD)**			
Accuracy	**Precision**	**Recall**	**AUC**	**Accuracy**	**Precision**	**Recall**	**AUC**
96.77%	**100%**	**90.91%**	**0.959**	95.16%	95.24%	90.91%	0.972

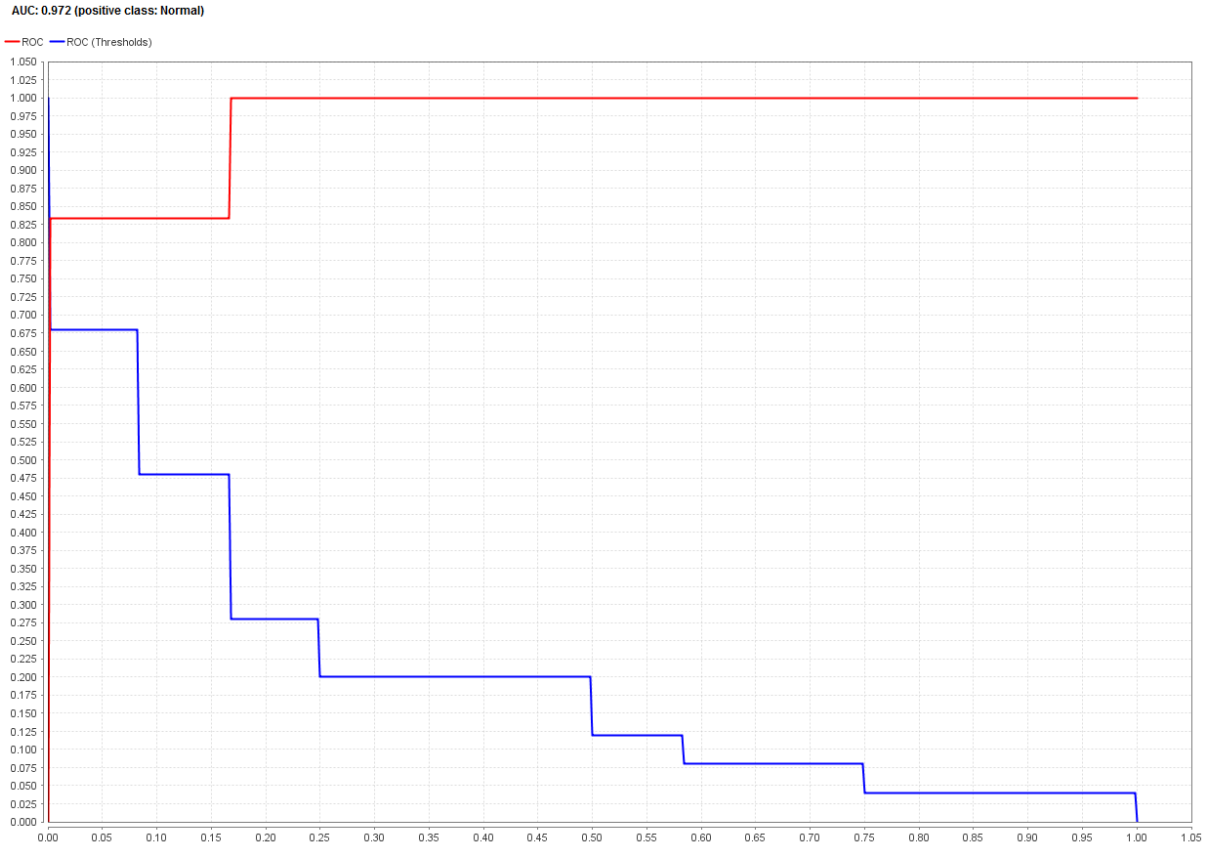

Figure 5.6-3: Colon Cancer Dataset ROC-AUC

5.6.2 Case 2: Prostate Cancer Dataset

The prostate cancer dataset [66] contains 12533 microarray gene expressions against 102 patients, among which 50 are with normal tissues, while 52 are with cancerous tissues.

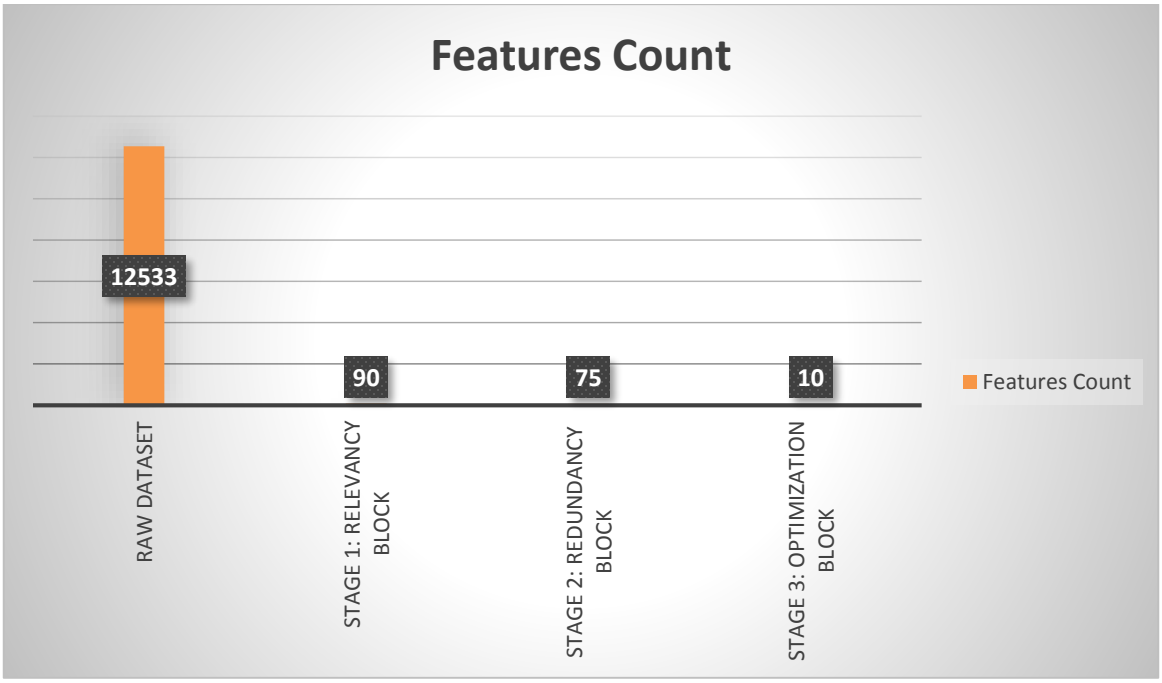

Figure 5.6-4: Prostate Cancer Dataset: Feature (Genes) Count after each stage

The figure 5.6-4 shows the count of selected features after each stage. Raw Prostate cancer dataset contains 12533 features (genes). In Relevancy Block, Information Gain Filter selected 55 Features, Gain Ratio filter selected 78 Features and Gini Index filter selected 68 Features. Union of Features by each filter meeting a specific threshold criteria resulted in a set comprising of 90 informative and relevant genes. In next stage, redundancy among features is removed and we removed 15 genes that were redundant and not adding any new information to the target class, resulting in 75 features.

These 75 genes are passed as a reduced search space for Genetic Algorithm wrapper, where fitness of each individual is calculated by the classification accuracy of Random Forest in a 10 fold stratified Cross Validation Setup. And we got a final optimal feature subset containing 10 potential features shown in table 5.6-3.

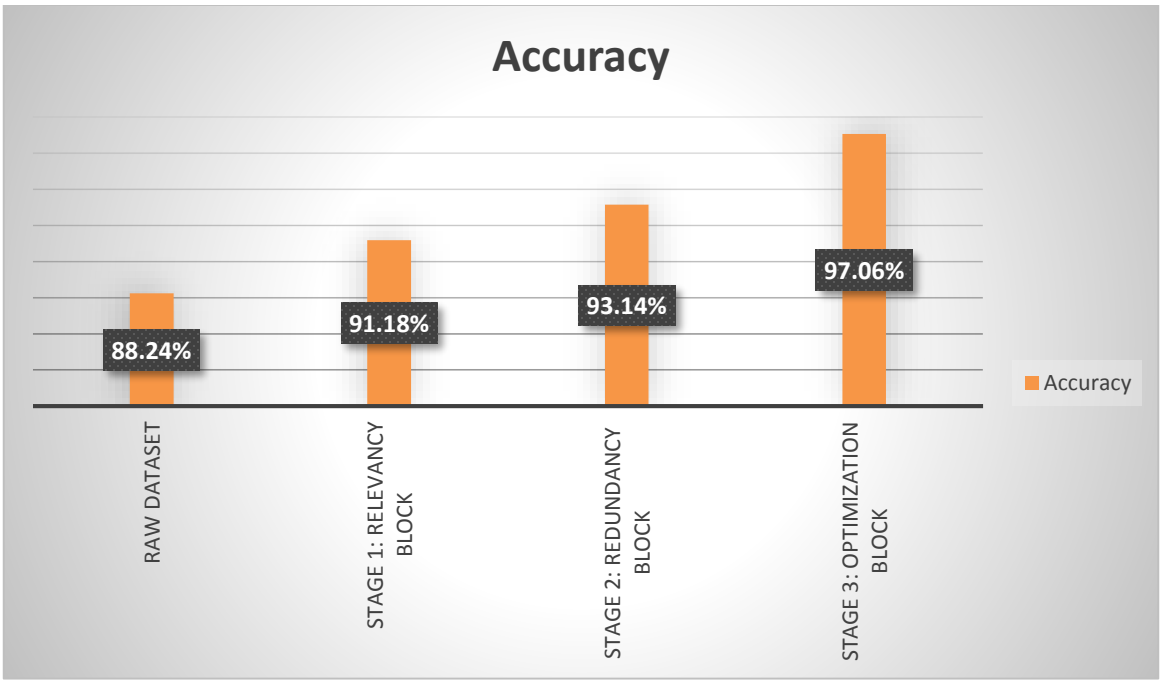

Figure 5.6-5: Prostate Cancer Dataset: Classification Accuracy after each stage

Figure 5.6-5 shows classification accuracy achieved by features set obtained at the end of each stage using random forest classifier in 10 fold stratified Cross Validation setup.

To create a baseline classification accuracy, we computed accuracy with Raw Dataset without feature selection i.e. 88.24%. And to witness impact of each stage on classification performance of classifier, we have computed classification accuracies at the end of each stage. We got 91.18% accuracy for prostate cancer dataset after stage 1 while 93.14% after stage 2, which clearly shows, Pearson correlation statistics not only removed redundant features but noise too, thus improving classification accuracy. Finally, the optimal feature set shown in table 5.6-3 has attained 97.06% classification accuracy.

Table 5.6-3: Prostate Cancer Dataset: Selected Optimal Feature Subset

Selected Potential Attributes	Accuracy Achieved
{863g_at, 1740_g_at, 1767_s_at, 33396_at, 36569_at, 37639_at, 38028_at, 38322_at, 39939_at, 41381_at}	97.06%

Table 5.6-4 shows accuracy, precision, recall and AUC measure for the validation split as well as for the test split. To avoid any chances of error and to assure the classification performance of selected feature subset, we have performed 10-fold Cross Validation, using different classifiers including Random Forest, SVM, Naïve Bayes and kNN. Here in the table 5.6-4 the best classification accuracy is shared which we have got with random forest Classifier.

Table 5.6-4: Prostate Cancer Dataset: Confusion Matrix

VALIDATION SPLIT (CV 10 FOLDS)				**TEST SPLIT**			
Accuracy	**Precision**	**Recall**	**AUC**	**Accuracy**	**Precision**	**Recall**	**AUC**
98.04%	**98.08%**	**98.08%**	**0.975**	97.06%	98.04%	96.15%	0.983

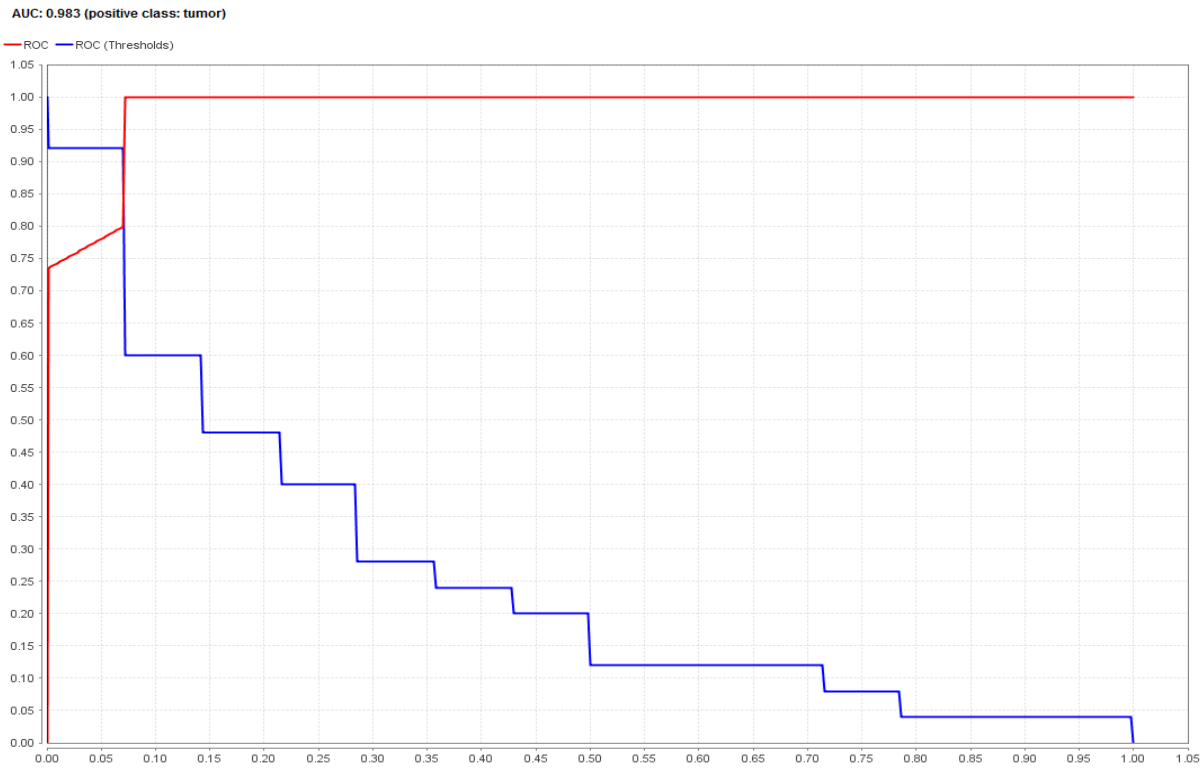

Figure 5.6-6: Prostate Cancer Dataset: ROC-AUC

5.6.3 Case 3: Leukemia Cancer Dataset

The Leukemia Cancer dataset is based on Leukemia cancer study by "*Golub et. al., (1999)*" [69]. Its contains 7129 microarray expression genes (features) profiled against 72 patients representing dataset instances, in which 49 instances corresponds to Acute Lymphoblast Leukemia (ALL) and 23 sample corresponds to Acute Myeloid Leukemia (AML).

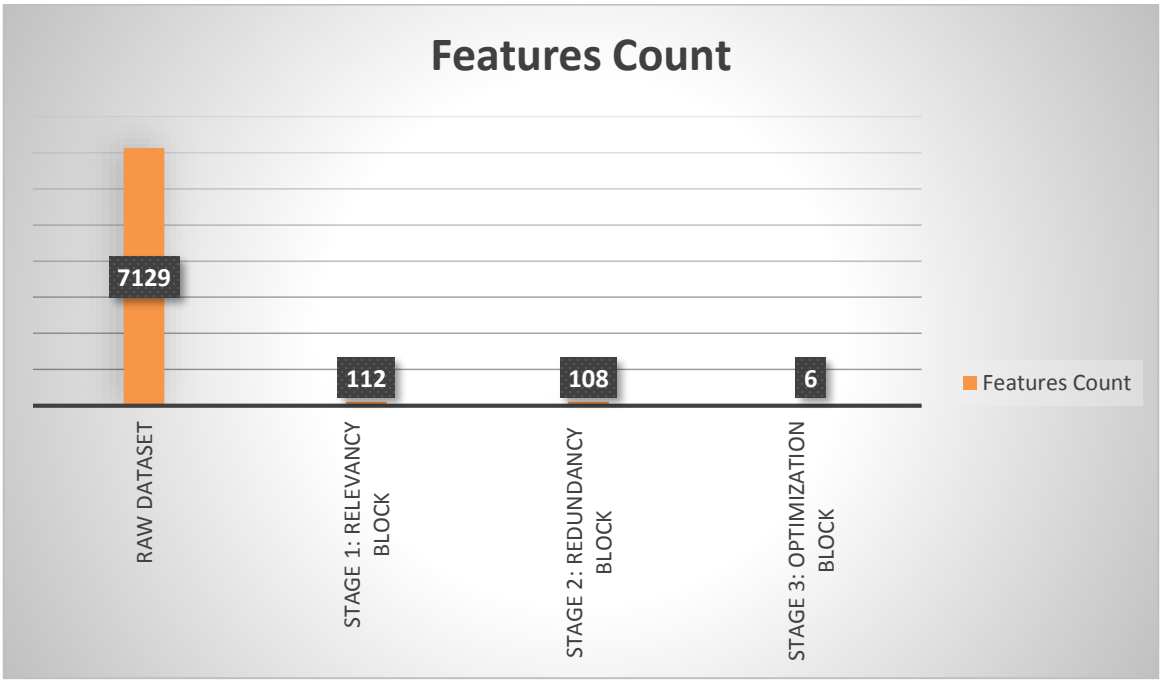

Figure 5.6-7: Leukemia Cancer Dataset: Feature (Genes) Count after each stage

The figure 5.6-7 shows the count of selected features after each stage. Raw Leukemia cancer dataset contains 7129 features (genes). In Relevancy Block, Information Gain Filter selected 64 Features, Gain Ratio filter selected 64 Features and Gini Index filter selected 87 Features. Union of Features by each filter meeting a specific threshold criteria resulted in a set comprising of 112 informative and relevant genes. In next stage, redundancy among features is removed and we removed 4 genes that were highly correlated and not adding any new information to the target class, resulting in 108 features.

These 108 genes are passed as a reduced search space for Genetic Algorithm wrapper, where fitness of each individual is calculated by the classification accuracy of Random Forest in a 10 fold stratified Cross Validation Setup. And we got a final optimal feature subset containing 6 potential features shown in table 5.6-5.

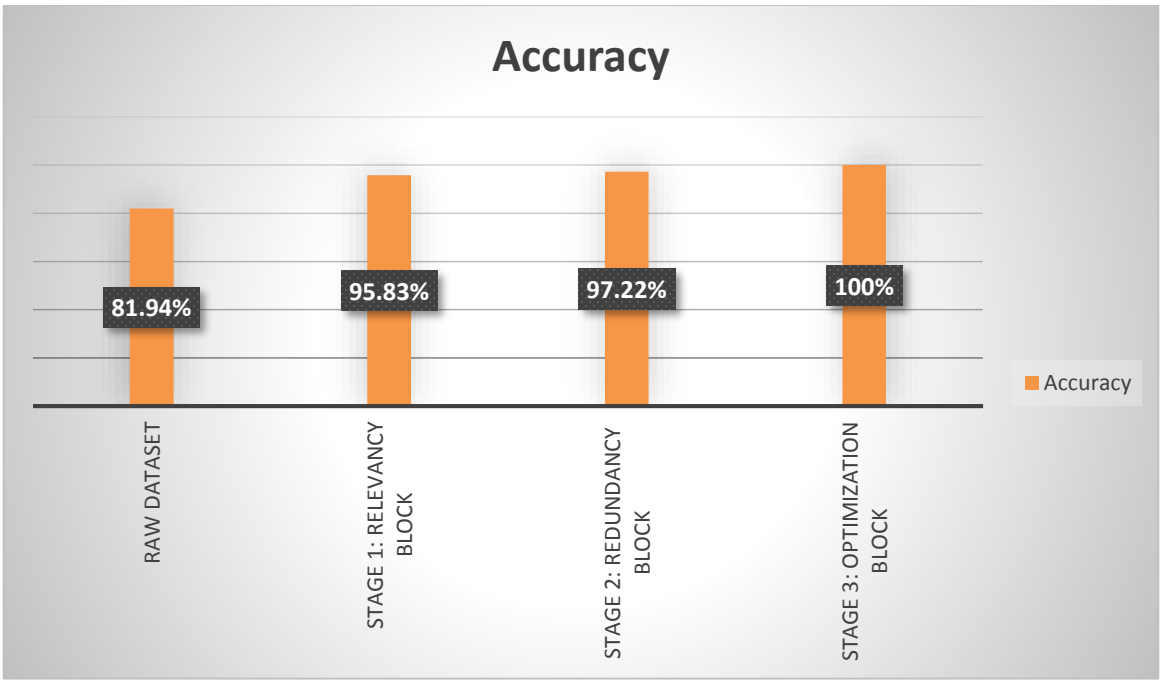

Figure 5.6-8: Leukemia Cancer Dataset: Classification Accuracy after each stage

Figure 5.6-8 shows classification accuracy achieved by features set obtained at the end of each stage using random forest classifier in 10 fold stratified Cross Validation setup.

To create a baseline classification accuracy, we computed accuracy with Raw Dataset without feature selection i.e. 81.94%. And to witness impact of each stage on classification performance of classifier, we have computed classification accuracies at the end of each stage. We got 95.83% accuracy for prostate cancer dataset after stage 1 while 97.22% after stage 2, which clearly shows, Pearson correlation statistics not only removed redundant features but noise too, thus improving classification accuracy. Finally, the optimal feature set shown in table 5.6-5 has attained 100% classification accuracy.

Table 5.6-5: Leukemia Cancer Dataset: Optimal Feature Subset

Selected Potential Attributes	Accuracy Achieved
{M89957_at, U16954_at, Y07604_at, M12959_S_AT, U29175_at, U46751_at}	100%

Table 5.6-5 shows the accuracy, precision, recall and AUC measure for the validation split as well as for the test split. To avoid any chances of error and to assure the classification performance of selected feature subset, we have performed 10-fold Cross Validation, using different classifiers including Random Forest, SVM, Naïve Bayes and kNN. Here in the table 5.6-6 the best classification accuracy is shared which we have got with random forest and Naïve Bayes Classifiers i.e.100%.

Table 5.6-6: Leukemia Cancer Dataset: Confusion Matrix

VALIDATION SPLIT (CV 10 FOLDS)				**TEST SPLIT (CV 10 FOLDS)**			
Accuracy	**Precision**	**Recall**	**AUC**	**Accuracy**	**Precision**	**Recall**	**AUC**
100	**100**	**100**	**1.0**	100	100	100	1.0

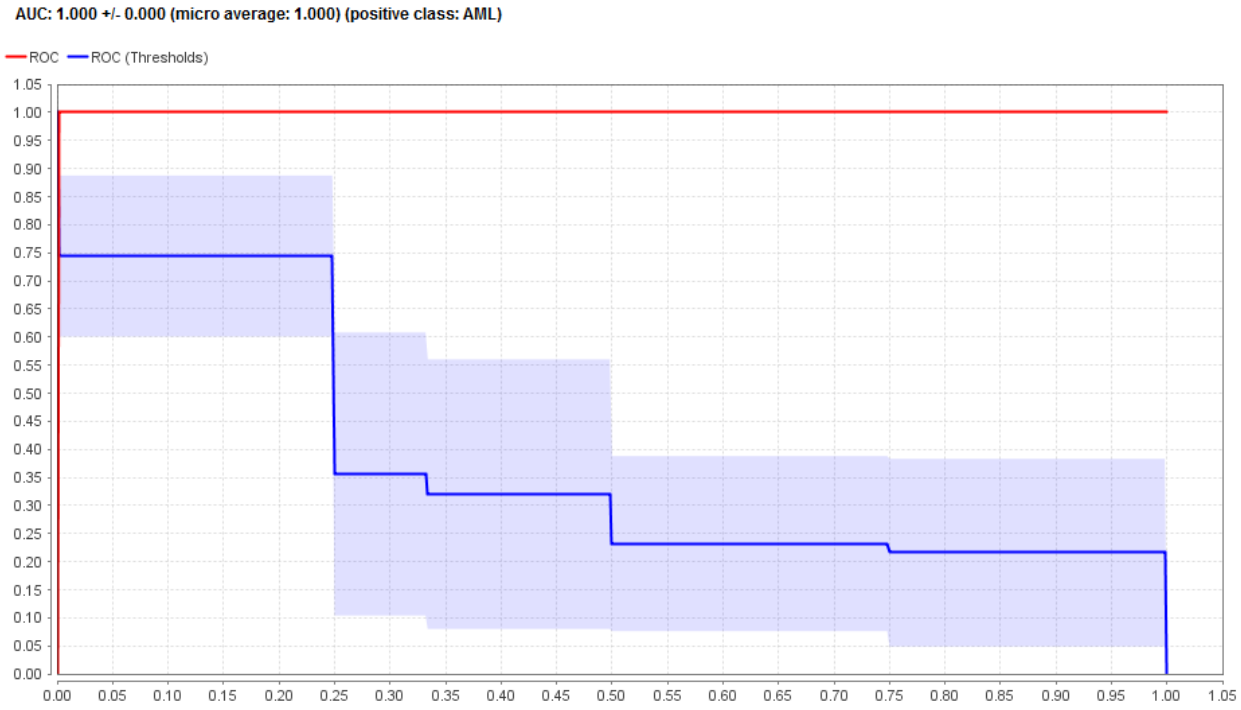

Figure 5.6-9: Leukemia Cancer Dataset: ROC-AUC

5.6.4 Case 4: Ovarian Cancer Dataset

The Ovarian Cancer Dataset [65] contains 15155 Microarray gene expressions (features) profiled against 253 patients (instances). Among these instances, 91 corresponds to Class "Normal", while 162 correspond to class "Cancer".

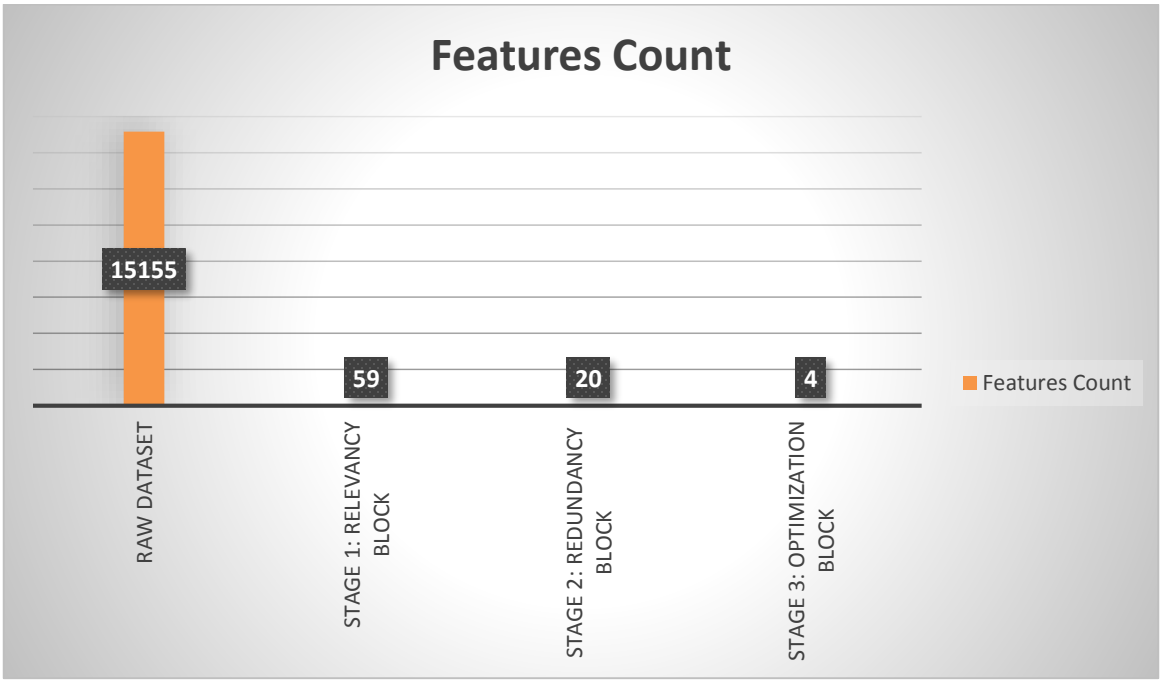

Figure 5.6-10: Ovarian Cancer Dataset: Feature (Genes) Count after each stage

The figure 5.6-10 shows the count of selected features after each stage. Raw Ovarian cancer dataset contains 15155 features (genes). In Relevancy Block, Information Gain Filter selected 42 Features, Gain Ratio filter selected 40 Features and Gini Index filter selected 57 Features. Union of Features by each filter meeting a specific threshold criteria resulted in a set comprising of 59 informative and relevant genes. In next stage, redundancy among features is removed and we removed 39 genes that were redundant and not adding any new information to the target class, resulting in 20 features.

These 20 genes are passed as a reduced search space for Genetic Algorithm wrapper, where fitness of each individual is calculated by the classification accuracy of Random Forest in a 10 fold stratified Cross Validation Setup. And we got two final optimal feature subset containing 4 potential features shown in table 5.6-7.

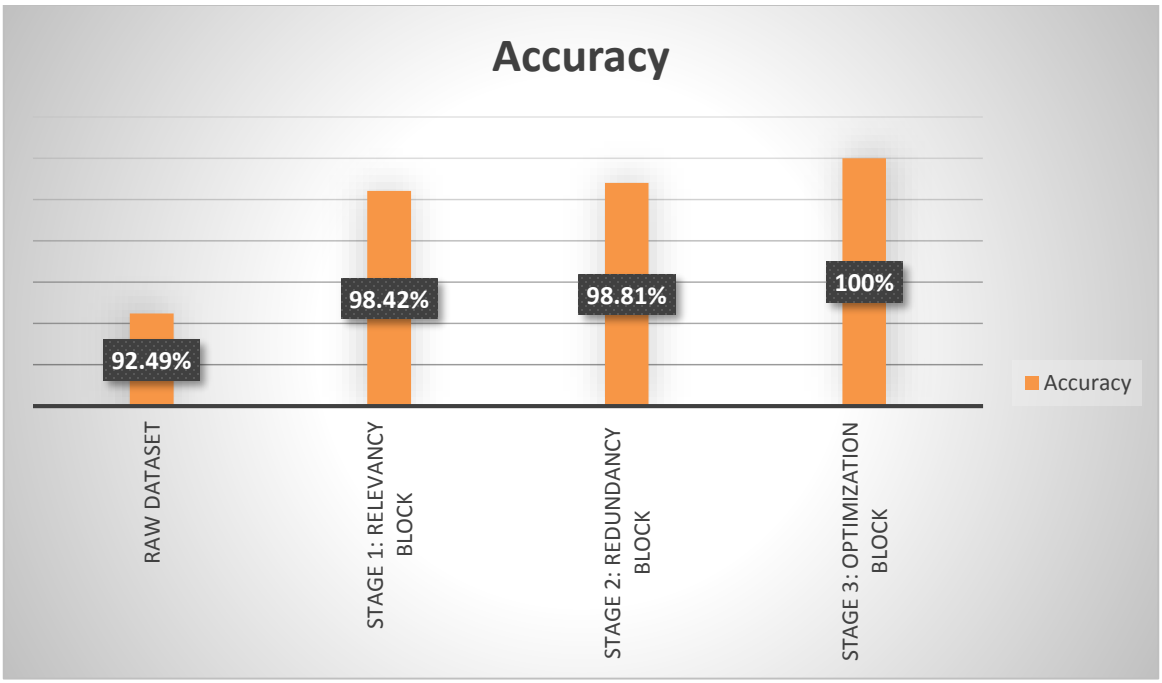

Figure 5.6-11:Ovarian Cancer Dataset: Classification Accuracy after each stage

Figure 5.6-11 shows classification accuracy achieved by features set obtained at the end of each stage using random forest classifier in 10 fold stratified Cross Validation setup.

To create a baseline classification accuracy, we computed accuracy with Raw Dataset without feature selection i.e. 92.49%. And to witness impact of each stage on classification performance of classifier, we have computed classification accuracies at the end of each stage. We got 98.42% accuracy for ovarian cancer dataset after stage 1 while 98.81% after stage 2, which clearly shows, Pearson correlation statistics not only removed redundant features but noise too, thus improving classification accuracy. Finally, the optimal feature set shown in table 5.6-7 has attained 100.00% classification accuracy.

Table 5.6-7: Ovarian Cancer Dataset: Selected optimal feature subset

Selected Potential Attributes	Accuracy Achieved
{MZ2.7921478, MZ244.07686, MZ247.295, MZ435.46452}	100%
{MZ2.7921478, MZ262.18857, MZ417.73207, MZ435.46452}	100%

Table 5.6-8 shows accuracy, precision, recall and AUC measure for the validation split as well as for the test split. To avoid any chances of error and to assure the classification performance of selected feature subset, we have performed 10-fold Cross Validation, using different classifiers including Random Forest, SVM, Naïve Bayes and kNN. Here in the table 5.6-8 the best classification accuracy is shared which we have got with kNN classifier i.e.100%.

Table 5.6-8: Ovarian Cancer Dataset: Confusion Matrix

VALIDATION SPLIT (CV 10 FOLDS)				**TEST SPLIT**			
Accuracy	**Precision**	**Recall**	**AUC**	**Accuracy**	**Precision**	**Recall**	**AUC**
100%	**100%**	**100%**	**1.0**	100%	100%	100%	1.0

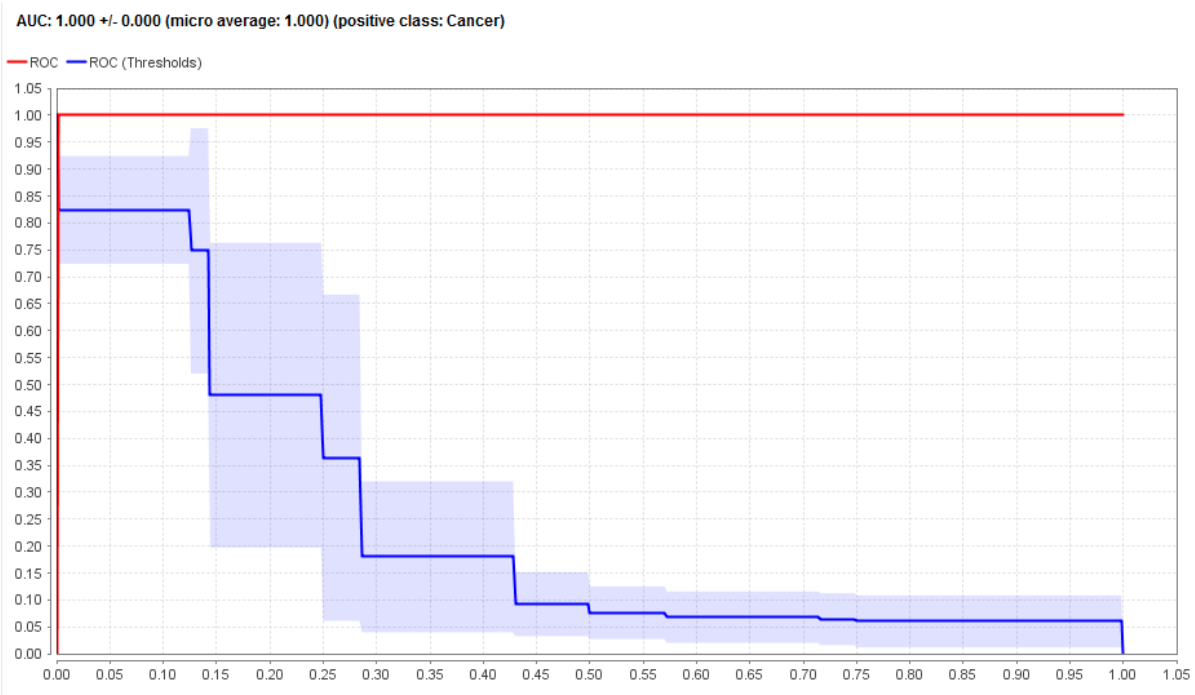

Figure 5.6-12: Ovarian Cancer Dataset: ROC-AUC

5.6.5 Case 5: Central Nervous System (CNS) Cancer Dataset

The Central Nervous System (CNS) Cancer Dataset [65] contains 7129 Microarray gene expressions (features) profiled against 60 patients (instances). Among these instances, 39 corresponds to Class "1" representing failures, while 21 correspond to class "0" representing survivors.

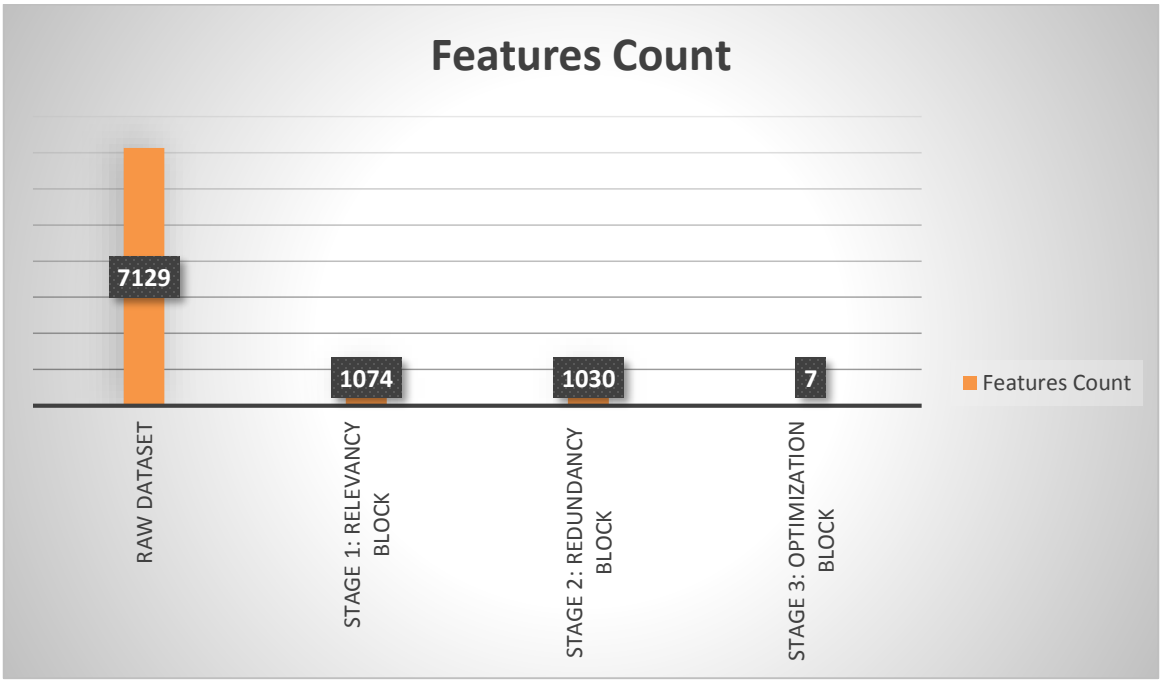

Figure 5.6-13: CNS Cancer Dataset: Feature (Genes) Count after each stage

The figure 5.6-13 shows the count of selected features after each stage. Raw CNS cancer dataset contains 7129 features (genes). In Relevancy Block, Information Gain Filter selected 88 Features, Gain Ratio filter selected 1015 Features and Gini Index filter selected 1015 Features. Union of Features by each filter meeting a specific threshold criteria resulted in a set comprising of 1074 informative and relevant genes. In next stage, redundancy among features is removed and we removed 44 genes that were redundant and not adding any new information to the target class, resulting in 1030 features.

These 1030 genes are passed as a reduced search space for Genetic Algorithm wrapper, where fitness of each individual is calculated by the classification accuracy of Random Forest in a 10 fold stratified Cross Validation Setup. And we got a final optimal feature subset containing 7 potential features shown in table 5.6-9.

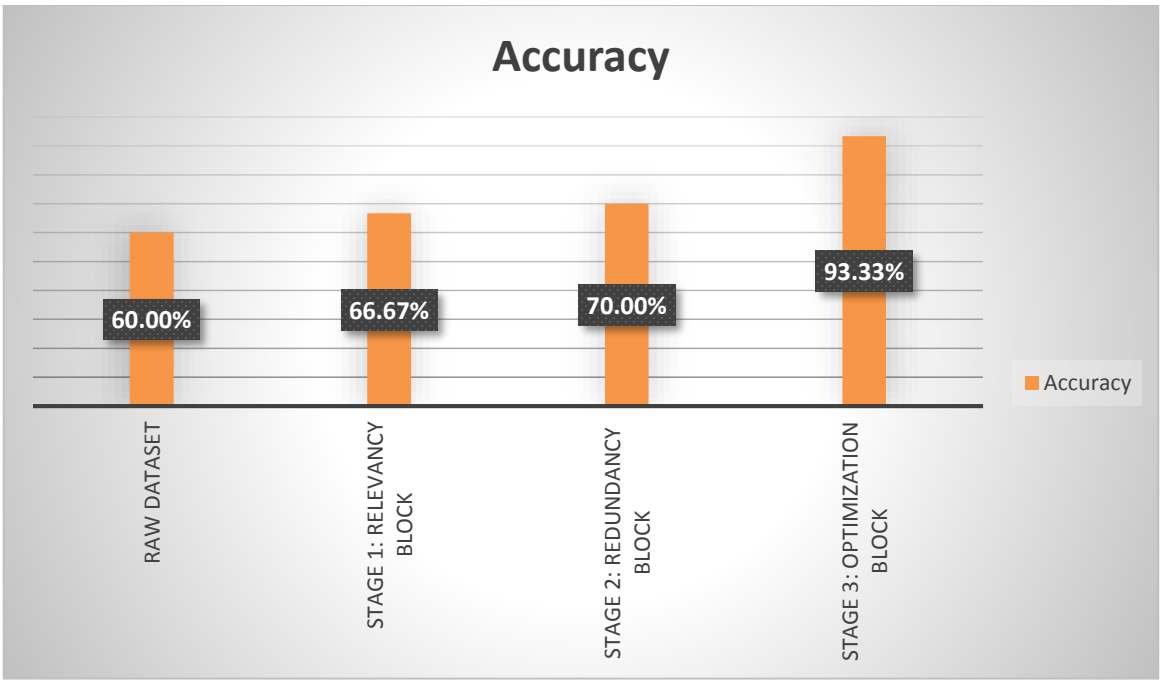

Figure 5.6-14: CNS Cancer Dataset: Classification Accuracy after each stage

Figure 5.6-14 shows classification accuracy achieved by features set obtained at the end of each stage using random forest classifier in 10 fold stratified Cross Validation setup.

To create a baseline classification accuracy, we computed accuracy with Raw Dataset without feature selection i.e. 60.00%. And to witness impact of each stage on classification performance of classifier, we have computed classification accuracies at the end of each stage. We got 66.67% accuracy for CNS cancer dataset after stage 1 while 70.00% after stage 2, which clearly shows, Pearson correlation statistics not only removed redundant features but noise too, thus improving classification accuracy. Finally, the optimal feature set shown in table 5.6-9 has attained 93.33% classification accuracy.

Table 5.6-9: CNS Cancer Dataset: Selected Optimal Feature Subset

Selected Potential Attributes	Accuracy Achieved
{D83542_at, S71824_at, U93205_at, X14968_at, D13814_s_at, U11821_s_at, X71348_at}	93.33%

Table 5.6-10 shows the accuracy, precision, recall and AUC measure for the validation split as well as for the test split. To avoid any chances of error and to assure the classification performance of selected feature subset, we have performed 10-fold Cross Validation, using different classifiers including Random Forest, SVM, Naïve Bayes and kNN. Here in the table 5.6-10 the best classification accuracy is shared which we have got with Random Forest classifier i.e. 93.33%.

Table 5.6-10: CNS Cancer Dataset: Confusion Matrix

VALIDATION SPLIT (CV 10 FOLDS)				**TEST SPLIT**			
Accuracy	**Precision**	**Recall**	**AUC**	**Accuracy**	**Precision**	**Recall**	**AUC**
93.33%	**94.87%**	**94.87%**	**0.912**	93.33%	92.68%	97.44%	.931

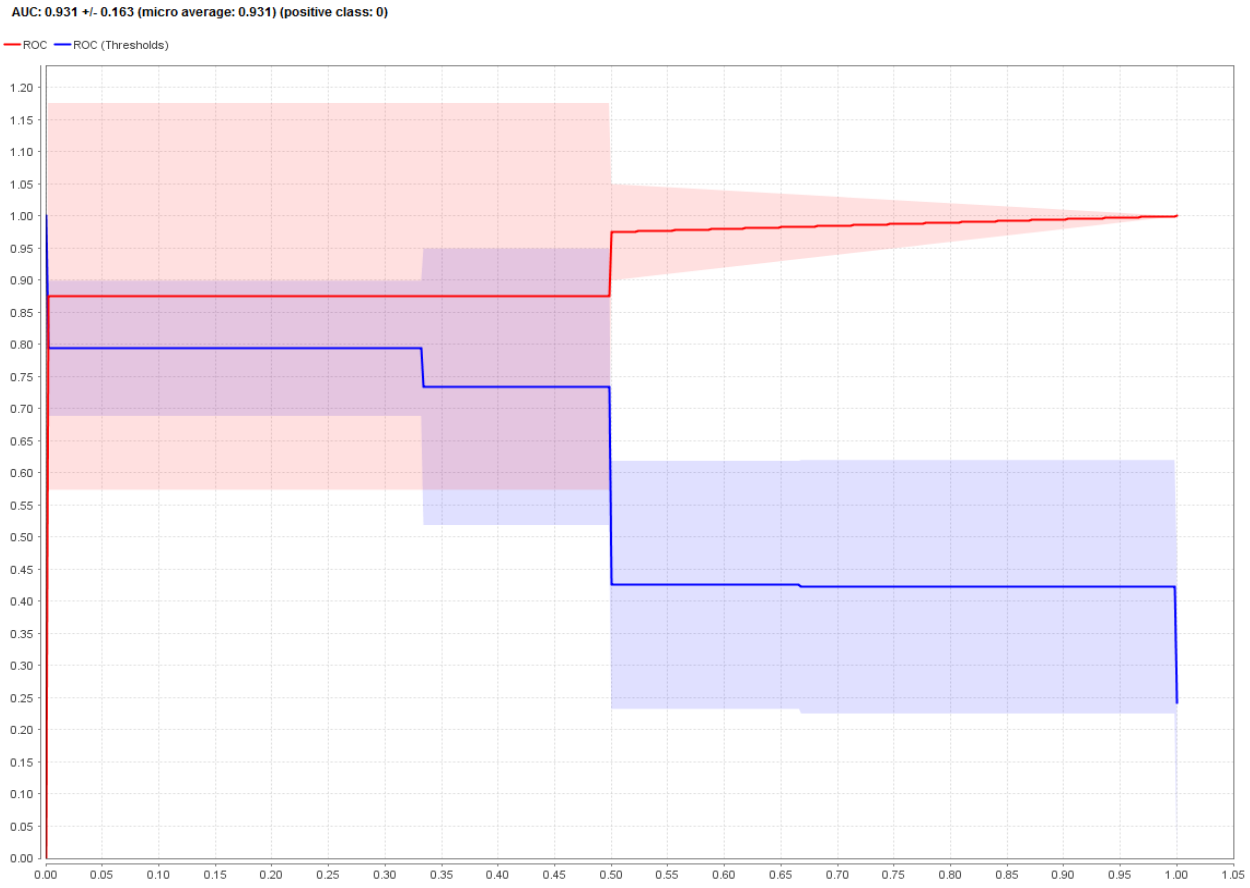

Figure 5.6-15- CNS Cancer Dataset: ROC-AUC

5.6.6 Case 6: Small Round Blue Cell Tumor (SRBCT) Dataset

The Small Round Blue Cell Tumor (SRBCT) dataset [67] contains 2308 microarray gene expression dataset profiled against 83 patients. The instances belong to four classes 1,2,3 or 4.

1, 2, 3 and 4 corresponds to Ewing family of tumors (EWS), non-Hodgkin lymphoma (NHL), neuroblastoma (NB), rhabdomyosarcoma (RMS) containing 29, 11, 18, 25 instances respect.

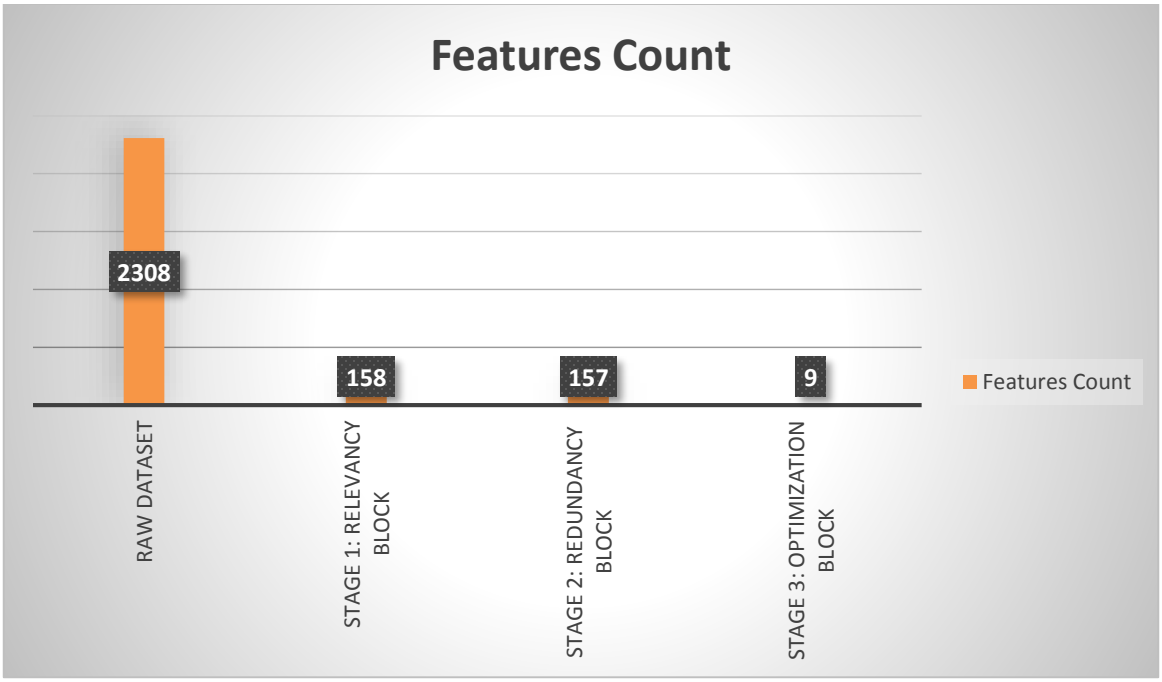

Figure 5.6-16: SRBCT Dataset: Feature (Genes) Count after each stage

The figure 5.6-16 shows the count of selected features after each stage. Raw SRBCT cancer dataset contains 2308 features (genes). In Relevancy Block, Information Gain Filter selected 121 Features, Gain Ratio filter selected 108 Features and Gini Index filter selected 63 Features. Union of Features by each filter meeting a specific threshold criteria resulted in a set comprising of 158 informative and relevant genes. In next stage, redundancy among features is removed and we removed 1 gene that was redundant and not adding any new information to the target class, resulting in 157 features.

These 157 genes are passed as a reduced search space for Genetic Algorithm wrapper, where fitness of each individual is calculated by the classification accuracy of Random Forest in

a 10 fold stratified Cross Validation Setup. And we got a final optimal feature subset containing 7 potential features shown in table 5.6-12.

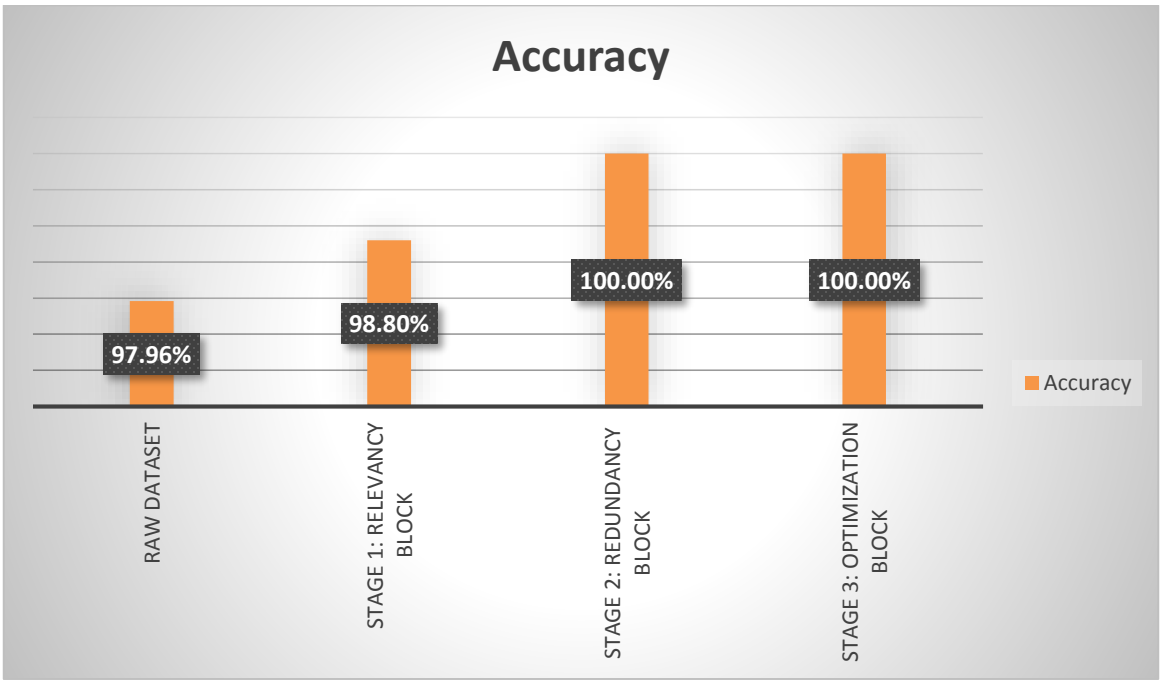

Figure 5.6-17: SSRBCRBCT Dataset: Classification Accuracy after each stage

Figure 5.6-17 shows classification accuracy achieved by features set obtained at the end of each stage using random forest classifier in 10 fold stratified Cross Validation setup.

To create a baseline classification accuracy, we computed accuracy with Raw Dataset without feature selection i.e. 97.96%. And to witness impact of each stage on classification performance of classifier, we have computed classification accuracies at the end of each stage. We got 98.80% accuracy for SRBCT multiclass cancer dataset after stage 1 and 100.00% for stage 2, which clearly shows, Pearson correlation statistics not only removed redundant features but noise too, thus improving classification accuracy.

Table 5.6-11: SRBCT Dataset: Selected Optimal Feature Subset

Selected Potential Attributes	Accuracy Achieved
{gene2, gene246, gene742, gene842, gene846, gene1764, gene 1770, gene1911, gene 2050}	100.00%

Table 5.6-12 shows accuracy, precision, recall and AUC measure for the validation split as well as for the test split. To avoid any chances of error and to assure the classification performance of selected feature subset, we have performed 10-fold Cross Validation, using different classifiers including Random Forest, SVM, Naïve Bayes and kNN. Here in the table 5.6-12 the best classification accuracy is shared which we have got with Naïve Bayes and Random Forest classifier i.e. 100.00%.

Table 5.6-12: SRBCT Dataset: Confusion Matrix

VALIDATION SPLIT (CV 10 FOLDS)			**TEST SPLIT (CV 10 FOLDS)**		
Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
100%	**100%**	**100%**	100%	100%	100%

5.6.7 Case 7: Lymphoma Cancer Dataset

The Lymphoma Cancer Dataset [65] is a multiclass dataset that contains 4026 microarray gene expression (features) profiled against 66 patients (instances). 46 instances correspond to Diffuse Large B-cell Lymphoma (DLBCL), 9 to Follicular Lymphoma (FL), 11 to Chronic lymphocytic leukemia (CLL). All three categories are cancer types that grow in white blood cells.

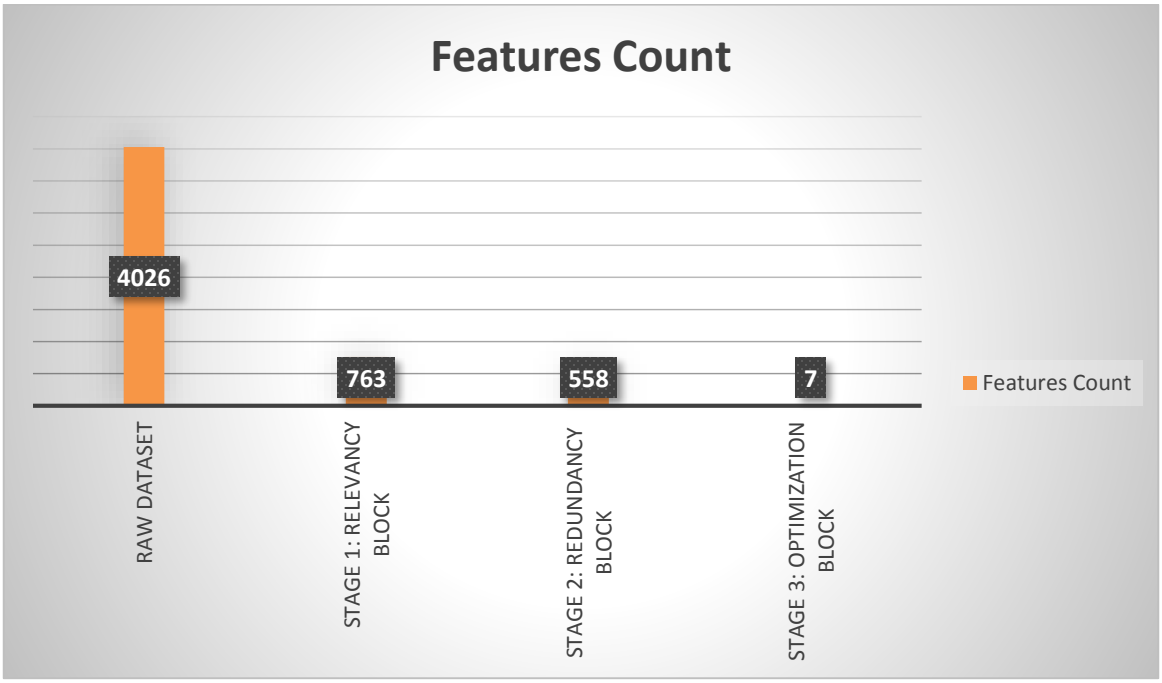

Figure 5.6-18: Lymphoma Cancer Dataset: Feature (Genes) Count after each stage

The figure 5.6-18 shows the count of selected features after each stage. Raw Lymphoma cancer dataset contains 4026 features (genes). In Relevancy Block, Information Gain Filter selected 402 Features, Gain Ratio filter selected 728 Features and Gini Index filter selected 504 Features. Union of Features by each filter meeting a specific threshold criteria resulted in a set comprising of 763 informative and relevant genes. In next stage, redundancy among features is removed and we removed 205 genes that was redundant and not adding any new information to the target class, resulting in 558 features.

These 558 genes are passed as a reduced search space for Genetic Algorithm wrapper, where fitness of each individual is calculated by the classification accuracy of Random Forest in a 10 fold stratified Cross Validation Setup. And we got a final optimal feature subset containing 7 potential features shown in table 5.6-13.

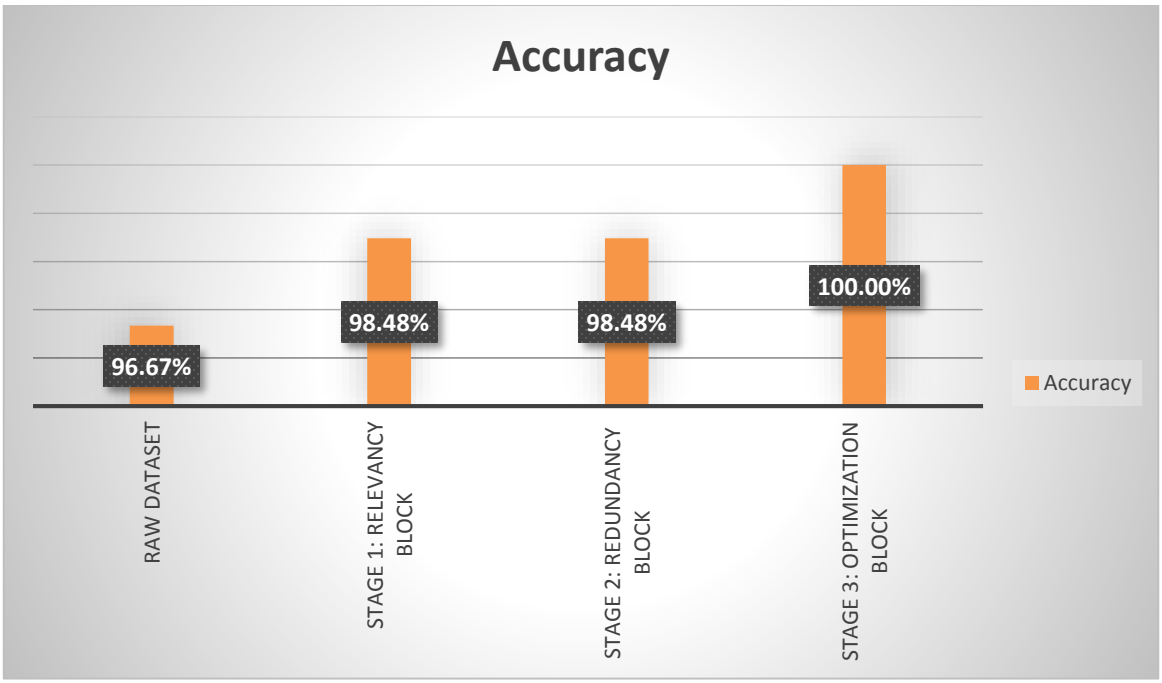

Figure 5.6-19: Lymphoma Cancer Dataset: Classification Accuracy after each stage

Figure 5.6-19 shows classification accuracy achieved by features set obtained at the end of each stage using random forest classifier in 10 fold stratified Cross Validation setup.

To create a baseline classification accuracy, we computed accuracy with Raw Dataset without feature selection i.e. 96.67%. And to witness impact of each stage on classification performance of classifier, we have computed classification accuracies at the end of each stage. We got 98.48% accuracy for Lymphoma multiclass cancer dataset after both stages, which clearly shows, Pearson correlation statistics has removed redundant features, just causing repetition and not adding any new information to target class. Finally, the optimal feature set shown in table 5.6-13 has attained 100.00% classification accuracy.

Table 5.6-13: Lymphoma Cancer Dataset: Selected Optimal Feature Subset

Selected Potential Features	Accuracy Achieved
{GENE894X, GENE1219X, GENE1672X, GENE1731X, GENE2080X, GENE2252X, GENE3594X}	100%

Table 5.6-14 shows the accuracy, precision, recall and AUC measure for the validation split as well as for the test split. To avoid any chances of error and to assure the classification performance of selected feature subset, we have performed 10-fold Cross Validation, using different classifiers including Random Forest, SVM, Naïve Bayes and kNN. Here in the table 5.6-14 the best classification accuracy is shared which we have got with Naïve Bayes and Random Forest classifier i.e. 100.00%.

Table 5.6-14: Lymphoma Cancer Dataset: Confusion Matrix

VALIDATION SPLIT (CV 10 FOLDS)			**TEST SPLIT (CV 10 FOLDS)**		
Accuracy	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
100%	**100%**	**100%**	100%	100%	100%

5.7 Comparative Analysis

In this section we have covered different set of comparisons. First comparison we have done is between GA-RF wrapper and MF-GARF Hybrid approach in terms of classification accuracy, features count and CPU execution time to show the performance gap.

Table 5.7-1: Comparison of GA-RF Wrapper and MF-GARF Wrapper
(Features Count is in parenthesis)

	DATASETS	**MF-GA-RF**	**Execution Time (sec)**	**GA-RF**	**Execution Time (sec)**
BINARY CLASS DATASETS	COLON	95.16 (5)	17s	91.94(10)	387s
	PROSTATE	97.06 (10)	280s	88.24(10)	1709s
	LEUKAEMIA	100 (6)	25s	97.22(10)	466s
	OVARIAN	100 (4)	121s	98.42(10)	1726s
	CNS	93.33 (7)	338s	88.33(8)	726s
MULTI CLASS DATASET	SRBCT	100 (9)	30s	95.18(10)	1068s
	LYMPHOMA	100 (7)	20s	98.48(10)	1547s

This table 5.7-1 presents a comparison among proposed hybrid approach MFGARF and wrapper approach GA-RF. The purpose is to show how preprocessing through filters impacts positively on classification accuracies and CPU execution time.

5.7.1 Comparison of our Proposed Multiple Filter Based Preprocessing with mRMR

This table 5.7.2 presents a comparison of our chosen set of filters with famous filter technique mRMR employed by many studies [45,48,49] to proof it's a good competitor and it

has achieved better accuracies for five datasets than mRMR. To keep the comparison fair we have considered same count of features for both methods.

Table 5.7-2: Comparison of accuracies achieved by Proposed Multiple Filters and mRMR
(Features count is represented in parenthesis)

	DATASETS	MULTIPLE FILTER (MF) FEATURE SET	mRMR FEATURE SET
BINARY CLASS DATASETS	COLON	87.10(42)	87.10(42)
	PROSTATE	93.14(75)	94.12(75)
	LEUKAEMIA	98.81(102)	97.22(102)
	OVARIAN	100(16)	98.42(16)
	CNS	70.00(1030)	71.67(1030)
CLASS DATASE	SRBCT	100(157)	97.59(157)
	LYMPHOMA	98.48(558)	98.48(558)

5.7.2 Comparison of GA-RF with commonly used GA-SVM combination

In table 5.7-3 we have made comparison of our proposed GA-RF with commonly used combination of GA-SVM in terms of accuracies and features count. To keep the comparison fair, Initial preprocessing is kept same for both combination i.e. proposed multiple filters, to ensure relevancy and remove redundancy, are employed in earlier stages. The average of accuracies shows that GA-RF is a better choice than GA-SVM.

Table 5.7-3: Comparison of Accuracies achieved by MFGARF and MF-GASVM
(Features count is represented in parenthesis)

		PROPOSED MULTIPLE FILTERS (MF)	
	DATASETS	GA-RF	GA-SVM
BINARY CLASS DATASETS	COLON	95.16(5)	94.44(10)
	PROSTATE	97.06(10)	83.10(10)
	LEUKAEMIA	100(6)	100(9)
	OVARIAN	100 (4)	100(6)
	CNS	93.33(7)	66.67(8)

MULTI CLASS DATASETS	SRBCT	100(9)	94.12(10)
	LYMPHOMA	100(7)	84.62(10)
	Average Accuracy	97.93%	88.99%

In table 5.7 2, we can see MF-GARF has achieved classification accuracies and features count better than MF-GASVM for 5 datasets and same accuracy for 2 datasets but with lesser number of features comparatively.

5.7.3 Comparison of MF-GARF with Other state of the art GA Wrapper based Hybrid Approaches

Here in table 5.7-4 we have presented a comparison of our proposed MFGARF Hybrid with other state of the art Genetic Algorithm based hybrid feature selection approaches in terms of classification accuracy and feature (genes) count. MFGARF approach has achieved 100 percent accuracy for Leukemia, Ovarian, SRBCT, and Lymphoma dataset with no more than 4-9 features which is not yet achieved by any of other GA based Feature selection hybrid approach. For colon, it has achieved better accuracy than IG-GA and MIMGA and FCBF – GA with relatively very small number of features i.e. 5, while other techniques feature count ranges between 14-202 for Colon Cancer Dataset. For prostate, we have outperformed L_Score-GA and F_Score-GA. MFGARF has outperformed IG-GA for CNS dataset too by achieving 93.3% accuracy with 7 features which is far better than 86.67% off 38 features.

Table 5.7-4: Comparison of MFGARF with other state of the art GA based Hybrid Approaches (Features count is represented in parenthesis)

Feature Selection Algorithm	**Classifier**	**Colon**	**Prostate**	**Leukemia**	**Ovarian**	**CNS**	**SRBCT**	**Lymphoma**
(IG-GA) [41]	GP	85.48 (60)	-	97.06(3)	-	86.67 (38)	-	-

(MIMAGA) [42]	ELM	89.09 (7)	96.54(3)	97.62(19)	-	-	-	-
(IDGA – F) [43]	SVM	-	96.8(25)	98.1(12)	-	-	97.6(21)	98.1(10)
	kNN	-	92.5(29)	98.1(14)	-	-	97.8(20)	95.4(8)
	NB	-	92.3(30)	95.2(18)	-	-	97.9(30)	97.9(9)
(IDGA – L) [43]	SVM	-	89.3(32)	95.6(13)	-	-	97.3(32)	99.7(21)
	kNN	-	73.6(33)	97.4(8)	-	-	96.9(29)	91.2(23)
	NB	-	56.7(25)	97.7(8)	-	-	96.1(29)	93.0(20)
SCC-GA [44]	SVM	68.75 (NAN)	-	-	-	-	89.02 (NAN)	-
	kNN	79.38 (NAN)	-	-	-	-	82.36 (NAN)	-
	NB	83.80 (NAN)	-	-	-	-	85.00 (NAN)	-
	DT	85.24 (NAN)	-	-	-	-	87.02 (NAN)	-
mRMR-GA [45]	SVM	**95.61** (83)	-	93.05(51)	-	-	-	-
mRMR-GA [48]	SVM	85.48 (NAN)	-	98.61 (NAN)	-	-	-	95.83 (NAN)
Extended GA [46]	ELM	-	-	97.4(124)	-	-	-	-
(FCBF – GA) [47]	SVM	90.32 (14)	-	-	100 (30)	-	-	-
Proposed MF-GARF	RF	95.16 (5)	**97.06 (10)**	**100(6)**	**100(4)**	**93.33 (7)**	**100 (9)**	**100 (7)**
	kNN	79.03 (5)	93.14 (10)	94.44 (6)	**100(4)**	60.00 (7)	91.57 (9)	93.94 (7)
	SVM	82.14 (5)	96.08 (10)	97.22 (6)	99.60(4)	68.33 (7)	-	-
	NB	72.58 (5)	96.08 (10)	**100 (6)**	98.83(4)	61.67 (7)	**100 (9)**	**100 (7)**

Results show the improved performance of GA. Other than achieving best classification accuracies and minimum feature count with Proposed Approach (MF-GARF) and Random Forest Classifier, our objective was also to achieve relevancy, removal of redundancy, and achieve optimality. We have also covered few other gaps we encountered in GA approaches like not only performed experimentation on binary class but also on Multi-class dataset to show the effectiveness of our technique on multi class datasets. We have also shared the potential feature subset, which almost every GA based hybrid approach has neglected.

5.7.4 Comparison of MGARF with other state of the art Metaheuristic Hybrid Approaches

In recent years, swarm based intelligence base approach Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Artificial Bee Colony (ABC), PSO have gained much popularity in field of data mining for the purpose of gene selection because of it early convergence rate and easy interpretability. In this section we have done comparison of our proposed hybrid approach MF-GARF with PSO, ACO, ABC based hybrid approaches from recent literature [47-57].

Table 5.7-5: Comparison of MFGARF with other state of the art Metaheuristic Hybrid approaches(Features count is represented in parenthesis)

Feature Selection Algorithm	**Classifier**	**Colon**	**Prostate**	**Leukemia**	**Ovarian**	**CNS**	**SRBCT**	**Lymphoma**
(mRMR-PSO)[49]	SVM	87.10(401)	71.43 (8.2)	-	-	68.10 (11)	-	-
	kNN	85.48(44)	68.10 (10)	-	-	69.52 (9)	-	-
(mRMR-PSO)[48]	SVM	93.55 (78)	-	95.83 (53)	-	-	93.97 (68)	96.96 (82)
(HPSO-LS) [40]	1-NN	84.38 (60)	-	89.28 (100)	-	-	-	-
(FCBF-PSO) [47]	SVM	90.32 (14)	-	-	-	-	-	100 (30)

(CLACOFS) [50]	NB	-	**99.10 (7)**	97.60(6)	-	-	-	-
	kNN	-	**99.85 (15)**	95.95(4)	-	-	-	-
	SVM	-	**98.35 (14)**	95.95(3)	-	-	-	-
(MI - ASCO) [51]	Fuzzy Classifier	-	84.68 (NAN)	94.75 (NAN)	-	-	-	-
ReliefF-ACO – GS [52]	-	94.00(9)	89.20 (10)	95.80(18)	-	-	-	-
(ICA + ABC)[53]	NB	**98.14(16)**	**98.88(16)**	98.68(12)	-	-	-	97.33 (15)
(mRMR-ABC)[49]	SVM	87.10(354)	69.05 (29)	-	-	71.43 (8)	-	-
	kNN	85.48(113)	70.00 (10)	-	-	71.43 (9)	-	-
(RFR-BBHA)[54]	Bagging	91.93(4)	-	-	-	86.66 (2)	-	-
(SU-HSA)[55]	IB1	87.15(22)	-	99.53(23)	99.94(15)	-	-	-
	NB	87.53(9)	-	100(26)	99.65(12)	-	-	-
Log-GOA[56]	-	95(NAN)	-	94(NAN)	-	-	-	-
Proposed MF-GARF	RF	95.16 (5)	97.06 (10)	**100(6)**	**100(4)**	**93.33 (7)**	**100 (9)**	**100 (7)**
	kNN	79.03 (5)	93.14 (10)	94.44 (6)	**100(4)**	60.00 (7)	91.57 (9)	93.94 (7)
	SVM	82.14 (5)	96.08 (10)	97.22 (6)	99.60(4)	68.33 (7)	-	-
	NB	72.58 (5)	96.08 (10)	**100 (6)**	98.83(4)	61.67 (7)	**100 (9)**	**100 (7)**

The table 5.7-4 gives a thorough comparison of our proposed Feature Selection Hybrid Approach MFGARF with other state of the art techniques in terms of classification accuracy and features count. MFGARF has shown remarkable performance almost for all datasets. MFGARF has outperformed all other metaheuristic techniques in terms of classification accuracy and number of features in four datasets including Ovarian, Lymphoma, SRBCT and Leukemia. It has obtained 100% accuracy for these four dataset with features count no more than 9 that is exceptional. For CNS Cancer dataset, it has outperformed all three techniques [49, 54] with 93.33% classification accuracy and 7 features. For colon dataset, MFGARF has obtained 95.16% classification accuracy with Random Forest Classifier which is better than classification accuracies of 10 out of 11 techniques. ICA-ABC [53] has outperformed our proposed approach by achieving 98.14% accuracy with 15 features. For prostate cancer dataset, CLA-ACO [50] and ICA-ABC [53] has achieved better accuracies 99.85% and 98.84% respectively which is better than ours 97.06%. And the basic reason we got to figure out is the difference of features and sample size. All the techniques we compared with have feature set for prostate cancer dataset comprising of 12600 features with 136 instances while ours is 12533 features with 102 sample size. Otherwise, this table 5.7-4 gives us a clear idea that overall our proposed MF-GARF is a good competitor in terms of classification accuracy and predictive features (genes) it has yielded.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Conclusion

The claim of this thesis is that challenges associated with microarray cancer datasets specifically curse of dimensionality and irrelevancy can be resolved by performing supervised feature selection utilizing efficiency of filters and classification performance of metaheuristic wrapper approaches. For that purpose, we have proposed a hybrid feature selection approach (MF-GARF) based on multiple filters and Genetic Algorithm (GA) wrapper in combination with Random Forest (RF) classifier.

We have combined set of information theory based filters Information Gain, Gain Ratio, Gini index, and Correlation statistics that are responsible for removal of irrelevancy and redundancy among features. Moreover, these multiple filters not only collectively serve as a pre-processor for GA wrapper but also improve the performance of GA by reducing the initial search space of GA wrapper that is otherwise highly computationally expensive. Genetic Algorithm and Random Forest with Gini Index Splitting criteria wrapper refines the pre-processed feature set into an optimal subset. For the evaluation of selected features Random Forest, kNN, Naïve Bayesian and SVM classifiers are used in 10-folds cross validation to avoid any chances of biasness and overfitting and we got the best results with Random Forest.

The experimentation has been carried out on 7 benchmark binary and multi-class microarray cancer datasets including Colon, Prostate, Leukemia, Ovarian, CNS, SRBCT, and Lymphoma. The proposed algorithm has achieved 100% accuracy for Leukemia, Ovarian, SRBCT and Lymphoma with 6, 4, 9, and 7 features respectively. For Colon we got 95.16% classification accuracy with 5 features and for prostate 97.06% with 10 features. We have compared the results of MFGARF with other state of the art GA wrapper based hybrid approaches and it has completely outperformed all other GA hybrid techniques for all datasets. Moreover, to evaluate it performance against other competitor metaheuristic approaches in terms of classification accuracy and features count we have compared MFGARF with PSO, ACO and ABC presented in literature. Our proposed approach has almost outperformed many hybrid techniques except few like CLACOFS and ICA-ABC for Colon and Prostate datasets.

6.2 Future Work

One of the aim of this study is to bring improvement in the performance of traditional Genetic Algorithm Wrapper, for which we have used fusion of multiple filters as a preprocessor and reduced the search space for GA wrapper. Moreover, we introduced Random Forest as an induction algorithm to evaluate the fitness of each selected feature (genes) subset in a 10-fold Cross Validation setup instead of SVM that is generally chosen as a fitness evaluator for GA Wrapper [38, 45, 47].

We successfully got the expected performance and results. But, we have left computational complexity as a future direction. Further, this work can either be refined or extended in many ways. For say, we can introduce novel Local Search algorithm with GA to improve the performance of GA wrapper as suggested by [40] [70] for PSO and BA respectively.

Here we have presented a comparison of our proposed hybrid approach with three other famous metaheuristic techniques namely Particle swarm optimization (PSO), Ant Colony Optimization (ACO) and Artificial Bee Colony Optimization (ABC) based hybrids. These metaheuristic techniques can be incorporated with proposed Preprocessing Step I.e. Fused Multiple Information theory based Filters.

For the purpose of noise reduction and removal of irrelevancy we have used set of information theory based filters information gain, Gain Ratio and Gini index, as they are considered as most promising ranking approaches [59]. For future work, set of filters with different feature evaluation criteria can be assembled, i.e. based on distance measures, similarity measures or consistency measures and their impact on classification performance can be studied.

Moreover, Feature selection of microarray gene expression datasets is a vast domain. There is still a great room to research, explore and try varied combination of existing algorithms to overcome the challenges and issues associated with microarray gene expression datasets and to improve the performance of existing Machine Learning Algorithms for such high dimensional datasets.

REFERENCES

[1] Li, J. and Liu, H., 2017. Challenges of feature selection for big data analytics. IEEE Intelligent Systems, 32(2), pp.9-15.

[2] Yang, X.S., Lee, S., Lee, S. and Theera-Umpon, N., 2015. Information analysis of high-dimensional data and applications. Mathematical Problems in Engineering, 2015.

[3] Rew, D.A., 2001. DNA microarray technology in cancer research. European Journal of Surgical Oncology, 27(5), pp.504-508.

[4] Berrar, D.P., Dubitzky, W. and Granzow, M. eds., 2003. A practical approach to microarray data analysis (pp. 15-19). New York: Kluwer academic publishers.

[5] Elizondo, D.A., Passow, B.N., Birkenhead, R. and Huemer, A., 2008. Dimensionality reduction and microarray data. In Principal Manifolds for Data Visualization and Dimension Reduction (pp. 293-308). Springer, Berlin, Heidelberg.

[6] Badaoui, F., Amar, A., Hassou, L.A., Zoglat, A. and Okou, C.G., 2017. Dimensionality reduction and class prediction algorithm with application to microarray Big Data. Journal of Big Data, 4(1), p.32.

[7] Cai, J., Luo, J., Wang, S. and Yang, S., 2018. Feature selection in machine learning: A new perspective. Neurocomputing, 300, pp.70-79.

[8] Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1), pp.16-28.

[9] F. He, H. Yang, Y. Miao and R. Louis, "A hybrid feature selection method based on genetic algorithm and information gain," 2016 5th International Conference on Computer Science and Network Technology (ICCSNT), Changchun, 2016, pp. 320-323.

[10] N. Prasad and M. M. Naidu, "Gain Ratio as Attribute Selection Measure in Elegant Decision Tree to Predict Precipitation," 2013 8th EUROSIM Congress on Modelling and Simulation, Cardiff, 2013, pp. 141-150.

[11] S. Sivagama Sundhari, "A knowledge discovery using decision tree by Gini coefficient," 2011 International Conference on Business, Engineering and Industrial Applications, Kuala Lumpur, 2011, pp. 232-235.

[12] Reggiani, Enrico, et al. "Pearson Correlation Coefficient acceleration for modeling and mapping of neural interconnections." 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). IEEE, 2017.

[13] Saidi, R., Ncir, W.B. and Essoussi, N., 2018, February. Feature Selection Using Genetic Algorithm for Big Data. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 352-361). Springer, Cham.

[14] Sehgal, S., Singh, H., Agarwal, M. and Bhasker, V., 2014, November. Data analysis using principal component analysis. In 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom) (pp. 45-48). IEEE.

[15] En.wikipedia.org. (2019). Latent semantic analysis. [online] Available at: https://en.wikipedia.org/wiki/Latent_semantic_analysis [Accessed 11 Jun. 2019].

[16] En.wikipedia.org. (2019). Linear discriminant analysis. [online] Available at: https://en.wikipedia.org/wiki/Linear_discriminant_analysis [Accessed 11 Jun. 2019].

[17] En.wikipedia.org. (2019). Independent component analysis. [online] Available at: https://en.wikipedia.org/wiki/Independent_component_analysis [Accessed 5 Dec. 2019].

[18] Muthukrishnan, R. and Rohini, R., 2016, October. LASSO: A feature selection technique in predictive modeling for machine learning. In 2016 IEEE International Conference on Advances in Computer Applications (ICACA) (pp. 18-20). IEEE.

[19] Yu, L. and Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 5(Oct), pp.1205-1224.

[20] Sánchez-Maroño, N., Alonso-Betanzos, A. and Tombilla-Sanromán, M., 2007, December. Filter methods for feature selection–a comparative study. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 178-187). Springer, Berlin, Heidelberg.

[21] Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. Artificial intelligence, 97(1-2), pp.273-324.

[22] Lal T.N., Chapelle O., Weston J., Elisseeff A., 2006 Embedded Methods. In: Guyon I., Nikravesh M., Gunn S., Zadeh L.A. (eds) Feature Extraction. Studies in Fuzziness and Soft Computing, vol 207. Springer, Berlin, Heidelberg

[23] Min, H. and Fangfang, W., 2010, December. Filter-wrapper hybrid method on feature selection. In 2010 Second WRI Global Congress on Intelligent Systems (Vol. 3, pp. 98-101). IEEE.

[24] Shen, Q., Diao, R. and Su, P., 2012. Feature Selection Ensemble. Turing-100, 10, pp.289-306.

[25] Cang, S., 2011, October. A mutual information based feature selection algorithm. In 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI) (Vol. 4, pp. 2241-2245). IEEE.

[26] Jin, X., Xu, A., Bie, R. and Guo, P., 2006, April. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In International Workshop on Data Mining for Biomedical Applications (pp. 106-115). Springer, Berlin, Heidelberg.

[27] Momenzadeh, M., Sehhati, M. and Rabbani, H., 2019. A novel feature selection method for microarray data classification based on hidden Markov model. Journal of biomedical informatics, 95, p.103213.

[28] Alirezanejad, M., Enayatifar, R., Motameni, H. and Nematzadeh, H., 2019. Heuristic filter feature selection methods for medical datasets. Genomics.

[29] Nagpal, A. and Singh, V., 2018. A feature selection algorithm based on qualitative mutual information for cancer microarray data. Procedia computer science, 132, pp.244-252.

[30] Dash, R., 2017. A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study. Journal of King Saud University-Computer and Information Sciences.

[31] Yuan, M., Yang, Z. and Ji, G., 2019. Partial maximum correlation information: A new feature selection method for microarray data classification. Neurocomputing, 323, pp.231-243.

[32] Hasan, A. and Adnan, M.A., 2012, February. High dimensional microarray data classification using correlation based feature selection. In 2012 International conference on biomedical engineering (ICoBE) (pp. 319-321). IEEE.

[33] Mandal, M. and Mukhopadhyay, A., 2013. An improved minimum redundancy maximum relevance approach for feature selection in gene expression data. Procedia Technology, 10, pp.20-27.

[34] Mishra, S. and Mishra, D., 2016. Enhanced gene ranking approaches using modified trace ratio algorithm for gene expression data. Informatics in Medicine Unlocked, 5, pp.39-51.

[35] Alshamlan, H.M., Badr, G.H. and Alohali, Y.A., 2016. Abc-svm: artificial bee colony and svm method for microarray gene selection and multi class cancer classification. Int. J. Mach. Learn. Comput, 6(3), p.184.

[36] Almugren, N. and Alshamlan, H., 2019, July. FF-SVM: New FireFly-based gene selection algorithm for microarray cancer classification. In 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) (pp. 1-6). IEEE.

[37] Vijay, S.A.A. and GaneshKumar, P., 2018. Fuzzy expert system based on a novel hybrid stem cell (HSC) algorithm for classification of micro array data. Journal of medical systems, 42(4), p.61.

[38] Hernandez, J.C.H., Duval, B. and Hao, J.K., 2007, April. A genetic embedded approach for gene selection and classification of microarray data. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (pp. 90-101). Springer, Berlin, Heidelberg.

[39] Chuang, L.Y., Yang, C.H., Li, J.C. and Yang, C.H., 2012. A hybrid BPSO-CGA approach for gene selection and classification of microarray data. Journal of Computational Biology, 19(1), pp.68-82.

[40] Moradi, P. and Gholampour, M., 2016. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Applied Soft Computing, 43, pp.117-130.

[41] Salem, H., Attiya, G. and El-Fishawy, N., 2017. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 50, pp.124-134.

[42] Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y. and Gao, Z., 2017. A hybrid feature selection algorithm for gene expression data classification. Neurocomputing, 256, pp.56-62.

[43] Dashtban, M. and Balafar, M., 2017. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics, 109(2), pp.91-107.

[44] Singh, P., Shukla, A. and Vardhan, M., 2017, November. Hybrid approach for gene selection and classification using filter and genetic algorithm. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 832-837). IEEE.

[45] El Akadi, A., Amine, A., El Ouardighi, A. and Aboutajdine, D., 2011. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information Systems, 26(3), pp.487-500.

[46] Yan, K. and Lu, H., 2018, October. An Extended Genetic Algorithm Based Gene Selection Framework for Cancer Diagnosis. In 2018 9th International Conference on Information Technology in Medicine and Education (ITME) (pp. 43-47). IEEE.

[47] Gao, L., Ye, M. and Wu, C., 2017. Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony. Molecules, 22(12), p.2086.

[48] Alshamlan, H., Badr, G. and Alohali, Y., 2015. mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. Biomed research international, 2015.

[49] Mohamed, N.S., Zainudin, S. and Othman, Z.A., 2017. Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data. Expert Systems with Applications, 90, pp.224-231.

[50] Sharbaf, F.V., Mosafer, S. and Moattar, M.H., 2016. A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics, 107(6), pp.231-238.

[51] Vijay, S.A.A. and GaneshKumar, P., 2018. Fuzzy expert system based on a novel hybrid stem cell (HSC) algorithm for classification of micro array data. Journal of medical systems, 42(4), p.61.

[52] Sun, L., Kong, X., Xu, J., Zhai, R. and Zhang, S., 2019. A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification. Scientific Reports, 9(1), p.8978.

[53] Aziz, R., Verma, C.K. and Srivastava, N., 2017. A novel approach for dimension reduction of microarray. Computational biology and chemistry, 71, pp.161-169.

[54] Pashaei, E., Ozen, M. and Aydin, N., 2016, February. Gene selection and classification approach for microarray data based on Random Forest Ranking and BBHA. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 308-311). IEEE.

[55] Shreem, S.S., Abdullah, S. and Nazri, M.Z.A., 2016. Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm. International Journal of Systems Science, 47(6), pp.1312-1329.

[56] Tumuluru, P. and Ravi, B., 2017. GOA-based DBN: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification. Int. J. of Appl. Eng. Research, 12, pp.14218-14231.

[57] Alshamlan, H., Badr, G. and Alohali, Y. (2015). Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational Biology and Chemistry, 56, pp.49-60.

[58] Osama, A.A., Khader, A.T., Al-Betar, M.A. and Alyasseri, Z.A.A., 2018. A hybrid filter-wrapper gene selection method for cancer classification. In 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS) (pp. 32-37).

[59] Shukla, A.K., Tripathi, D., Reddy, B.R. and Chandramohan, D., 2019. A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges. Evolutionary Intelligence, pp.1-21.

[60] Coley, D.A., 1999. An introduction to genetic algorithms for scientists and engineers. World Scientific Publishing Company.

[61] En.wikipedia.org. (2019). Random forest. [online] Available at: https://en.wikipedia.org/wiki/Random_forest [Accessed 2 Sep. 2019].

[62] En.wikipedia.org. (2019). K-nearest neighbors algorithm. [online] Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm [Accessed 10 Oct. 2019].

[63] En.wikipedia.org. (2019). *Naive Bayes classifier*. [online] Available at: https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Accessed 10 Oct. 2019].

[64] En.wikipedia.org. (2019). *Support-vector machine*. [online] Available at: https://en.wikipedia.org/wiki/Support-vector_machine [Accessed 15 Oct. 2019].

[65] Csse.szu.edu.cn. (2019). Microarray Datasets. [online] Available at: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html [Accessed 5 Sep. 2019].

[66] Biolab.si. (2019). Bioinformatics Laboratory. [online] Available at: http://www.biolab.si/supp/bi-cancer/projections/info/prostata.html [Accessed 8 Sep. 2019].

[67] Biolab.si. (2019). Bioinformatics Laboratory. [online] Available at: http://www.biolab.si/supp/bi-cancer/projections/info/SRBCT.html [Accessed 6 Sep. 2019].

[68] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 96(12), pp.6745-6750.

[69] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. and Bloomfield, C.D., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), pp.531-537.

[70] Dashtban, M., Balafar, M. and Suravajhala, P. (2018). Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics, 110(1), pp.10-17.